VOLUME XX[A] DECEMBER, 1928 PARTS III AND IV

BIOMETRY AND CHRONOLOGY.

BY KARL PEARSON.

SIR ISAAC NEWTON appears to have been the first to attempt to make a scientific study of chronology and to apply the idea of probability to testing the legends of early history. His works on chronology were issued after his death by John Conduit, his executor and nephew-in-law, in a work entitled: *The Chronology of Ancient Kingdoms amended, to which is prefixed: A Short Chronicle from the First Memory of Things in Europe to the Conquest of Persia by Alexander the Great.* By Sir Isaac Newton, London, 1728. It is noteworthy that Newton's chronological writings, like his theological, are usually dismissed without careful examination and as a product of a disregardable old age*. Yet it is unusual for a great man even in old age to write absolutely idle things, and those who will take the trouble to study either Newton's theological or his historical views will rapidly be convinced how far ahead he was of the average mind of his age†. This being the centenary year of the publication of Newton's *Chronology*, it may not be amiss to give some account of it and the investigations which owe their origin to Newton's suggestions.

That Newton had endeavoured to reconstruct the current chronology was known before his death. A certain Abbé Conti, a man of noble Venetian family, had made the acquaintance of a friend of Newton's, when he was visiting England about 1718. To this friend Newton had communicated a manuscript chronological index, which Newton desired should be kept secret. The friend obtained permission to communicate this index to Conti, on the condition that the latter also kept it secret. Conti, however, having got the index, proceeded to disperse copies of it in France; he actually got it translated by a French antiquary, who without Newton's permission published it with his own confutations in 1725. Thus Newton in his old age was dragged once more into that public wrangling which was of all things most distasteful to him. Conti even assisted Leibniz in an attempt to engage Newton in fresh disputes. Newton replied to his translator and critic in a paper which appears in Vol. XXXIV of the *Philosophical Transactions*, pp. 315—321.

* Both the chronological and theological writings are only referred to in the Mathematical Association's recent *Isaac Newton, 1642—1727*, in the bibliography, and then as "Minor Works."

† Newton's rejection of the Doctrine of the Trinity and his grave doubts as to the Divinity of Christ were not based on the rationalism of modern scientific agnosticism. On the contrary he had studied the Bible in the original texts, and he was better versed than many theologians of his day in the history of the Early Church and the writings of the early Fathers. It was on these studies that he founded his criticism of the current dogmas; and because he did so, his theological writings were pronounced worthless, and have remained so characterised for two centuries.

There is no doubt that Newton was badly treated in the matter, not only in the publication without his consent of a private document, but also in the nature of the criticisms made on it, which were intended to further embroil Newton, who had questioned accepted chronology, with the theologians who had already begun to suspect his religious orthodoxy. We are not at present concerned with Newton's advanced theological standpoint, but solely with the fact that in his chronological index he had for the first time fixed the average length of a reign at 18 to 20 years. The critic confusing *generation* with *reign*, which is a different matter, proceeded to demolish Newton on the basis of the critic's own error. Newton in reply cites the first 24 Kings of France (from Pharamond on) as reigning 458 years, or on the average about 19 years apiece. The next 24 Kings (from Ludovicus Balbus on) reigned 451 years, or 18¼ years apiece. The following 15 Kings (from Philippus Valesius on) reigned 315 years, or one with another 21 years apiece. Finally all the Kings of France, 63 in number, reigned 1224 years, or 19½ years apiece, or if we add in the long reign of Louis XIV, the 64 Kings reigned 20 years apiece. Newton adds:

"And they that examine the matter will find it so in other Kingdoms. And I shorten the duration of the antient Kingdoms of Greece, in the same proportion that I shorten the reigns of their Kings, and thereby place the Argonautic Expedition about 44 years and the Taking of Troy about 76 years after the death of Solomon, and find Sesostris contemporary to Sesac."

In Newton's *Short Chronicle*, which is prefixed to his *Chronology of Ancient Kingdoms amended*, he says (p. 7) that the 17 reigns of the Kings of Sparta take up 622 years, which is after the rate of 36½ years to a reign, and yet a Race of 17 Kings of that length is nowhere to be met with in true History and Kings at a moderate reckoning reign but 18 or 20 years apiece one with another. It is on this basis that Newton reconstructed ancient chronology and made it "fit better with the course of Nature."

On p. 51 of the main portion of his volume he tells us that the Egyptians, Greeks and Latins made reigns of Kings equipollent to generations. Thus he says the 7 Kings of Rome reigned 244 years, which is 35 years apiece. The first 12 Kings of Sicyon 529 years, or 44 years apiece. The first 8 Kings of Argos 371 years, or 46 years apiece. The 9 Kings of Acadia 379 years, or 42 years apiece; and so on. Thus the early historians gave 35 or 40 years to a reign. Newton says that on the other hand the 18 Kings of Judah who succeeded Solomon reigned 390 years, or 22 years apiece. The 15 Kings of Israel after Solomon reigned 259 years, or 17⅓ years apiece. The 18 Kings of Babylon 209 years, which is only 11½ years apiece. The 10 Kings of Persia 208 years, or 21 years apiece; and so on until Newton comes to the 30 Kings of England from William the Conqueror onwards with 648 years, or 21½ years apiece. Then Newton cites the data for the Kings of France we have already referred to and concludes that 18 to 20 years is the general average period for a reign. The fabulous periods provided by the early historians arise, he believes, from a confusion of generation and reign. He checks his reconstruction made in this manner by the records of the positions of the stars at the equinox and by other astronomical considerations. This brief account must suffice to show not only the

wide extent of Newton's historical reading, but also to indicate that he was the first man who strove to reconstruct chronology upon a scientific basis. We are not here concerned with his reconstruction of the individual chronologies of Egyptian, Persian, Syrian, Greek and other kingdoms, but only with his manner of approaching the subject, as the basis for later researches.

Voltaire refers in several of his writings to Newton's work. He confirms Newton's 18 to 20 years average reign on the reigns of the German Emperors. But he says that in states where revolutions are frequent, as in the Roman Empire from Caesar to Augustus (which gives 90 reigns in 500 years), Newton's rule shows too large a value*.

Voltaire attributes to the creative mind of Newton the discovery of a new method of settling chronological problems. For Newton in all he undertook did not fail to change accepted ideas; accustomed to disentangle chaos, he carried light into ancient fables confounded with history, and strove to reform an uncertain chronology. Voltaire writes thus:

"It is true there is no family, no town, no nation, which does not seek to throw back its origin. What is more the earliest historians are the most negligent in recording their dates. Books were a thousand times less common than to-day, and consequently less exposed to criticism; one could deceive the world with greater impunity; and while one misrepresented the facts, so one probably misrepresented the dates. In general it seemed to Newton that the world was 500 years younger than the Chronologists said it was. He based his idea on the ordinary course of Nature and upon astronomical observations†."

Voltaire, after saying that Newton's scheme appears to be correct, continues:

"Perhaps learned men will find it excessive to accord to the same man the honour of having perfected at once, physics, geometry and history. It would be a kind of universal monarchy to which their *amour-propre* would not easily accommodate itself. Thus at the time that the supporters of vortices and of porous matter attacked a demonstrated gravitation, Souciet and Fréret wrote against the chronology of Newton before it was published‡."

Voltaire later in his account of China applies Newton's rule to determine a probable duration of the reigns of the six Chinese Kings who preceded Hiao, and grants them 130 years, which he says will be more in accordance with the order of Nature than the 240 years attributed to the seven Kings of Rome and similar estimates contradicted by the experience of all ages§.

As a matter of fact mankind does not multiply itself as rapidly as some writers have assumed; and Voltaire in his day used Newton's chronology to demonstrate the absurdity of the enormous population of China being due to a repopulation of the world since the date then accepted of the Biblical Deluge.

It will be seen that Voltaire contributes little but approval to the method proposed by Newton. Condorcet on the other hand does, but his fault lies in suggesting that Voltaire, his beloved master, had contributed more to the science of chronology than his approval of Newton. The sixth section of Condorcet's

* *Oeuvres complètes*, t. XIII. p. 613 ("Annales de l'empire").

† *Oeuvres*, 1878, t. XXII p. 145, Lettre XVII. "Sur l'infini et sur la chronologie."

‡ *Loc. cit.* p. 148.

§ *Oeuvres*, t. XI. p. 466.

"Mémoire sur le Calcul des Probabilités" occupies pp. 454—468 of the *Mémoires de l'Académie Royale des Sciences* for 1784, published in 1787. As long as an occurrence can be treated as a discrete event, we can compute the probability of its occurrence. But if there are many alternative events the probability of a single event may be very small indeed. Accordingly nowadays, in estimating the improbability of a rare event, we are accustomed to calculate the probability of this event or of events still rarer. The commonest illustration is, perhaps, that in which we sum all the terms in the "tail" of a binomial in order to measure a probability "as rare or rarer." When the given events form a continuous series, so that the frequency of a *given* value of a quantity appearing can be represented only by a vanishingly small quantity, then we are accustomed to calculate from a frequency curve the chance of the given quantity lying between *two* definite values, for it is idle to compute the chance that it will take a definite value. In particular we integrate the tail of our frequency curve in estimating chances as rare or rarer than a given value. This process seems so familiar to us now that we find it difficult to place ourselves in the position of those who, in the time of Condorcet, wanted to determine whether a given event was so improbable that the authenticity of its occurrences must be questioned. It was a real puzzle to the writers on probability of those days, who could calculate the, often very small, chance of a given event, to proceed farther and get some measure of the general chance of such an event happening. Who first used the "as rare or rarer" formula I cannot at present say, but simple as it seems, it was a great advance. Condorcet in order to get over the difficulty introduces what he terms the "appropriate probability" ("la probabilité propre," pp. 454—455), but his ideas, to say nothing of his definitions, of this quantity appear to be very obscure. He sees that the probability b of the given event B does not cover what he has in mind, but he fails to define clearly his mental concept, or to obtain its algebraic representation.

In Section VII. p. 460, Condorcet gives a general expression for his "appropriate probability," not his only one, but one which is, perhaps, of all his attempts the easiest to follow. It runs thus: Suppose a to be the probability of A and b of B, $a > b$, and that there are m events A and n events B. Then Condorcet puts the mean of all the other probabilities other than that of a special b against b, i.e.

$$\frac{ma + (n-1)\,b}{m+n-1} \;:\; b,$$

and defines the "appropriate probability" of the event B as

$$\frac{b}{b + \dfrac{ma + (n-1)\,b}{m+n-1}} = \frac{(m+n-1)\,b}{ma + (m+2n-2)\,b}.$$

If $m = n$ and both be very large this equals $2b/(a+3b)$, and if a is nearly unity equals $2b$, or the "appropriate probability" is twice the ordinary probability. I confess to finding extreme difficulty in interpreting this. But taking his final result as $2b$ we may write it as

$$b + \tfrac{1}{2}b + \tfrac{1}{4}b + \tfrac{1}{8}b + \ldots,$$

or we might state that Condorcet is considering the events rarer than b to form a geometrical series of radix $\frac{1}{2}$. This interpretation has no root in Condorcet's disquisition, but it is the only way in which I can try to grasp what he is aiming at.

This preamble is necessary because when Condorcet comes to extend Newton's solution—which after all was an empirical one and gave no measure of inherent improbability—he is able to determine the chance that seven kings shall have reigned *exactly* 257 years*, but his only means of deducing from this the general improbability of the statement of the early historians is to apply to that value his extremely vague conception of "la probabilité propre."

On p. 461 of his memoir Condorcet says that the theory of probability has been applied to criticise and correct history and chronology, and that Newton led the way in this matter by applying our knowledge of the mean duration of generations and of reigns. Some philosophers indeed have used the theory of probability to question the veracity of historians†. He tells us that the savant Fréret had criticised such applications of the theory of probability and asserted that this theory should be confined to games of chance. Condorcet admits that certain writers have reached strange results by too hypothetical a use of the calculus, and by applying false principles, but he holds that Fréret was clearly not acquainted with the works of Halley, the Bernoullis and De Moivre. He refers to Voltaire as having applied the theory of probability to historical events, and in particular to the very long duration of the reigns of the seven Kings of Rome. I doubt very much whether Condorcet knew Newton's work at first hand, or he would have realised that Voltaire was really only repeating Newton's statements. Condorcet takes this classical case of the seven Kings of Rome reigning 257 years as an illustration of how the theory of probability may be applied to history. Now in this case it may not possibly be justifiable to apply the reigns of our own kings, who succeed according to a law of inheritance, because the Kings of Rome were elected. If a king lives to a great age, it will be likely that his successor, if a son, will have a short reign, and the influence may be felt even in the length of reign of his grandson. Further the kingship in an early state of society was not a very stable affair; kings were apt to be murdered, or killed in battle or deposed. Taking our own sovereigns from 1689 to 1902, and calling William and Mary one reign, we have eight rulers who died in their beds, and the exceptionally long reigns of George III and Victoria; the 213 years give us an average reign of 26·6 years, a long way from the 36·7 years apiece reported of the seven Kings of Rome. These latter were however *elected*. The only comparable material I could hit upon was formed by the German Emperors. From 800—1740, with a break of 19 years, there were

* Condorcet takes 257 years, not 244 years, as the total of the reigns of the seven Roman Kings, and I have adopted his value.

† Condorcet adds himself to their number by discussing on the basis of a somewhat specious theory the truth of the report that the Augur Accius Naevius, in order to convince Tarquin of his possession of miraculous powers, cut a flint with a razor.

only 50 of them and they give an average length of reign of 18·44 years, in striking accordance with Newton's result for non-elected kings*.

Condorcet does not approach the problem from experience as to the length of reign, but adopts what might be the better method, if we had the adequate data, namely experience as to the duration of life. He applies De Moivre's Law of Mortality to the universe of sovereigns—in this case to elected Roman Kings. According to that law, if N be the population and 90 years the limit of life, $N(90-A)$ are the number of persons alive at age A. Let m be the greatest number of years the oldest elected king can live and n the greatest number of years which the youngest elected king can live, then the number of persons from whom the king may be selected is

$$S_{90-n}^{90-m} N(90-y) = N(n+n-1+n-2+\ldots+m)$$
$$= \tfrac{1}{2}N(n+m)(n-m+1).$$

Now according to De Moivre's Law the number dying at age y in one year is $N(90-y)-N(90-y-1)=N$. Therefore the number dying in all the groups Nn, $N(n-1)$, $N(n-2)$,... is in each the same and equals N.

If we inquire as to how many die at the several ages, then:

If $s<m+1$, *all* the above groups are included and $N(n-m+1)$ will die.

If $s>n+1$, this excludes all the groups and none will die.

If s lies between $m+1$ and $n+1$, only those will die where there are persons with age s or greater, that is the groups from n down to s, i.e. $n-s+1$ groups, or $N(n-s+1)$ persons will die.

Now Condorcet wants to find a generative function, such that the coefficient of x^s therein shall take the above discontinuous values. He states that it is

$$N\left(\frac{(n-m+1)x}{1-x} - \frac{x^{m+1}-x^{n+2}}{(1-x)^2}\right),$$

but he does not state how he found this. It is easy to prove that it is and fulfils its purpose. We can show it thus:

If s be $<m+1$, we need only consider the first expression and the coefficient of x^s is clearly $N(n-m+1)$.

* It is not very satisfactory to base one's conclusions on 50 reigns, but I did work the problem out on this data in the first place in the manner in which I deal with more ample data later in this paper. The standard deviation was 13·258 years and the $\beta_1=\cdot6949$; it seemed idle to determine a β_2 on 50 cases. If we take the distribution of means of samples of seven we find: Mean of means of samples $=18\cdot44$, Standard Deviation $=5\cdot0111$ years, B_1 of curve of means $=\cdot0993$. This is not zero, but sufficiently small for us to assume the distribution of means to be approximately symmetrical, and (?) normal. We then have for the mean reign of the seven Roman Kings a deviation from the mean of means of $36\cdot71-18\cdot44=18\cdot27$, and its ratio to the standard deviation $5\cdot0111$ equals $3\cdot65$. This corresponds to $\frac{1}{2}(1+\alpha)=\cdot99987$, or there is only a probability of $\cdot00013$, of seven reigns having a length as long or longer than the reported reigns of the Kings of Rome. On this data we should accordingly be justified in holding the historians who report such reigns to be untrustworthy.

Now consider the second expression, we may write it in the form

$$N\frac{x^{m+1}-x^{n+2}}{(1-x)^2}=N\frac{x^{m+1}+x^{m+2}+\ldots+x^{n+1}}{1-x},$$

and, if s lies between $m+1$ and $n+1$, all the terms from x^{m+1} to x^s will contribute N to an x^s power, and the term is $s-(m+1)+1$, or the contribution $N(s-m)$. But the first term in the generative function still contributes $N(n-m+1)$; hence the total is

$$N(n-m+1)-N(s-m)=N(n-s+1).$$

Finally if s be $>n+1$, the power of the last term in the numerator of the second part, all the terms will come into play, or the total result will be

$$N(n-m+1)-N(n+1-m)=0.$$

Hence the probability of a person living to s years is the coefficient of x^s in

$$\frac{(n-m+1)x(1-x)-x^{m+1}+x^{n+2}}{\frac{1}{2}(1-x)^2(m+n)(n-m+1)}.$$

Or, if we raise this to the qth power and pick out the coefficient of x^{qs}, this will be the chance that q reigns will last qs years. Condorcet takes $n=60$ and $m=30$, or supposes no king would be elected younger than 30 or older than 60. For the Roman reigns $q=7$. The problem therefore turns on finding the coefficient of x^{257} in the expansion of

$$\left\{\frac{31x-31x^2-x^{31}+x^{62}}{45\times 31(1-x)^2}\right\}^7.$$

A modern statistician coming to the problem, would, I think, ask: What is the probability of the seven kings reigning 257 or more years? He would therefore have to sum the coefficients of x^{257} and *all higher powers.*

Condorcet does not proceed in this way. He finds the coefficient of x^{257} and says it is ·000792. I have not endeavoured to test his arithmetic, but shall be pleased if any of my readers will; more so still if they will find an approximation to the sum of the coefficients of all powers of x greater than 257! The convergence may be rapid, but I am doubtful of it*.

What Condorcet has really obtained is the chance of the seven reigns lasting *exactly* 257 years, and he does not know how from this to get a solution of his problem. What he does is to apply his "appropriate probability." He says the reigns may have had any duration from seven reigns of one year to seven reigns of 60 years, and he considers $413+1=414$ events possible. If $P=$ ·000792 Condorcet takes the mean probability of the other events to be $(1-P)/413$, and working on P and $(1-P)/413$ gets for his

$$\text{Appropriate Probability: } \frac{P}{P+\frac{1-P}{413}}=\frac{413P}{1+412P}=\cdot 2466=\frac{1}{4}, \text{ nearly,}$$

so that he makes odds of only 3 to 1!

* There are 24 terms in the numerator which do not surpass x^{257} and there are 257 further such terms in the expansion of $(1-x)^{-11}$. It does not therefore seem possible to evaluate all terms *less* than x^{257}.

The reader may easily see the weakness of Condorcet's "Probabilité propre," if he asks himself whether getting as many as seven heads in ten throws of a coin is to be considered an extraordinary event. The Condorcet method makes it more probable than not*.

It is clear, however, that our investigator got farther than Newton. He has proposed a method which depends upon a life-table, and, although we have no life-table of the Romans in the days of its early kings, we can use much later experience to provide at any rate an approximate solution. We can, I think, propound the problem in two different ways, neither of which is doubtless free from criticism, but if their numerical results converge we may hold the judgment based on them to be fairly satisfactory. In stating the problem in terms of the seven Kings of Rome we are not overlooking the fact that the methods of solution apply to a far wider range of problems.

1*st Method.* The frequency distribution of the reigns of sovereigns is given. What is the chance that a sample of seven selected at random will give a total length of reigns of 257 or more years?

2*nd Method.* Seven men are chosen at random between the ages of 30 and 60. Find the chance that their total future lifetime will be not less than 257 years.

The most forcible criticism of the first and easier method is that we have not the frequency distribution of reigns in the same country and at the same time as the reigns we are questioning. Anyone, however, who averages up reigns of different countries at different times will be surprised to find how closely the means are concentrated between 18 and 20 years, as Newton told us. And further how little difference there is between the reigns of hereditary and elected sovereigns. Once accept the distribution of reigns as approximately true, and the problem becomes a fairly easy one in the theory of small samples.

In the second method of approach to a solution there arise two main criticisms; the first is the choice of ages between which the king is to be selected. If we must determine d ages probably those adopted by Condorcet are as reasonable as most; although it is very possible that men of 30 and 60 would be less likely to be selected than men between 40 and 50. If the kingship is hereditary, then a real complexity is introduced by the negative correlation between adjacent reigns. The reigns cannot be taken "at random," nor the king selected at random from the available population. The second criticism of this method lies in the fact that we must choose a definite life-table to work from which will not be that of the time and country whose historians we are criticising. But just as the average reigns of

* My experience has been that, if a *small* term in a probability series be taken, those less than it can generally be represented approximately by a geometrical series. In this case we should have

$$P+rP+r^2P+\ldots=P/(1-r);$$

r can be determined if we compute *two* terms. If we give r its mean for all possible values, i.e. $\frac{1}{2}$, then $2P$ is our approximation. In this case we should find with Condorcet's value of P for our probability of 257 or more years the value ·0016. This is of the same order as the values found by the two methods discussed below.

kings do not seem to have changed much in duration till at any rate recent times, so we may not find between ages of 30 and 60 such changes in the life-tables as would prevent the second method giving a reasonable approximation to the desired frequency.

1*st Method.* With the assistance of Miss M. Child the length of reigns of 250 European sovereigns were extracted from various histories. It is, perhaps, worth noting that the Scandinavian Kings had the longest reigns, averaging about 20 years, while the elected sovereigns had, as we might expect, a lower average of about 18 years. Thus Newton's statement that average reigns run from about 18 to 20 years was amply comfirmed. The frequency distribution is as follows:

Central Values for Length of Reign.

Years	1·5	4·5	7·5	10·5	13·5	16·5	19·5	22·5	25·5	28·5	31·5	34·5	37·5	40·5	43·5	46·5	49·5	52·5	55·5	58·5	61·5	64·5	67·5	70·5	73·5	Total
Frequency	13	19	23	25	16	23	20	20	16	12	14	12	11	2	5	6	5	—	2	1	2	2	—	—	1	250

The constants of this distribution are:

Mean = 21·2520 years, Standard Deviation = 14·2693 years,

$\beta_1 = \cdot 803{,}247$, $\beta_2 = 3{\cdot}598{,}422$,

Modal Reign = 9·9583 years.

It is clear that the distribution is very skew.

I thought it theoretically best to fit the data with a Type I curve starting from zero length of reign, i.e. to use this start, the mean, standard deviation and β_1 to graduate the frequency. The resulting curve had for its equation

$$y = 23{\cdot}3076\left(1 - \frac{x}{92{\cdot}8442}\right)^{4{\cdot}9635}\left(1 + \frac{x}{10{\cdot}2583}\right)^{{\cdot}5484},$$

the origin being at the mode of 10·2583 years, which only differs by four months from the mode as found from both β_1 and β_2.

This curve and the histogram of observed reigns is shown in Diagram I.

Breaking into 15 groups we have:

Central Ages	1·5	4·5	7·5	10·5	13·5	16·5	19·5	22·5	25·5	28·5	31·5	34·5	37·5 — 40·5	43·5 — 46·5	49·5 — 73·5
Observed ...	13	19	23	25	16	23	20	20	16	12	14	12	13	11	13
Graduating Curve	13·4	21·2	23·5	23·8	22·9	21·4	19·5	17·4	15·4	13·3	11·4	9·7	14·7	9·7	12·6

Applying the goodness of fit test we have $\chi^2 = 4{\cdot}58$, and $P = \cdot 99$. Thus the graduation is satisfactory. We have now to consider the distributions of samples of seven drawn at random from this distribution of reigns. We know that:

Mean of means of samples = Mean of sampled population = 21·2520 yrs.

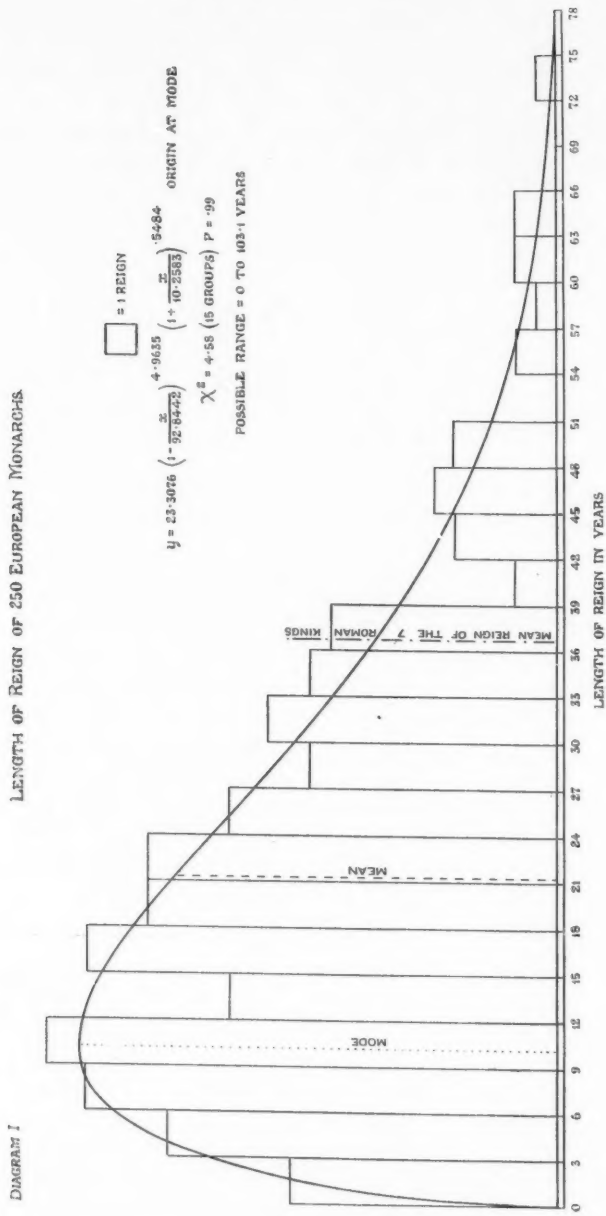

LENGTH OF REIGN OF 250 EUROPEAN MONARCHS
= 1 REIGN
$y = 23{\cdot}3076\left(1-\frac{x}{92{\cdot}8442}\right)^{4{\cdot}9655}\left(1+\frac{x}{10{\cdot}2585}\right)^{{\cdot}5484}$ ORIGIN AT MODE
$\chi^2 = 4{\cdot}58$ (15 GROUPS) $P = {\cdot}99$
POSSIBLE RANGE = 0 TO 103·1 YEARS
MODE
MEAN
MEAN REIGN OF THE 7 ROMAN KINGS
0 3 6 9 12 15 18 21 24 27 30 33 36 39 42 45 48 51 54 57 60 63 66 69 72 75 78
LENGTH OF REIGN IN YEARS

DIAGRAM I

Standard Deviation of means of samples $= \dfrac{\text{S.D. of sampled population}}{\sqrt{n}} = 5{\cdot}39328$ yrs.

$$B_1 = \frac{\beta_1}{n} = {\cdot}1147, \qquad B_2 = 3 + \frac{\beta_2 - 3}{n} = 3{\cdot}0855.$$

The curve distribution of means of samples of seven belongs therefore to Type I and its equation for a total frequency of 1000 is

$$y = 74{\cdot}1306 \left(1 + \frac{x}{22{\cdot}1247}\right)^{13{\cdot}2506} \left(1 - \frac{x}{88{\cdot}2722}\right)^{53{\cdot}0224}.$$

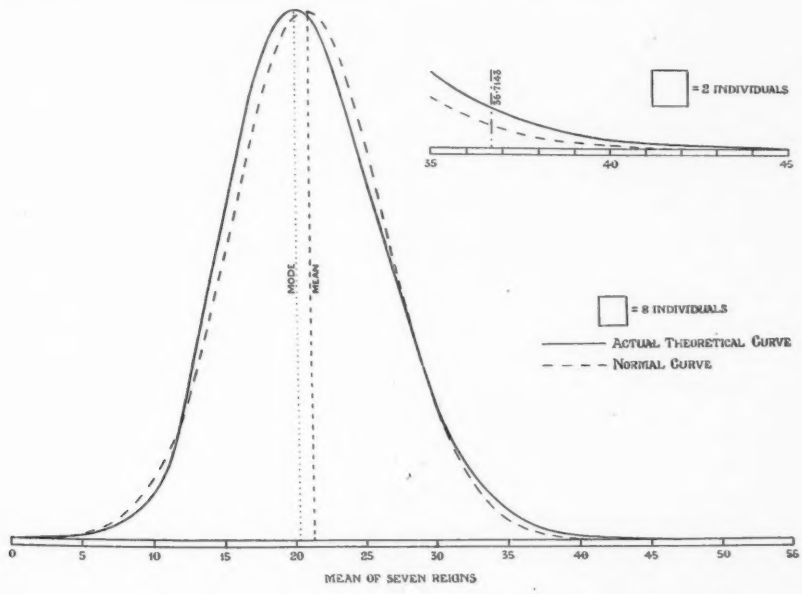
Origin at the mode 20·2837 years (Mean 21·2520 years). The possible range from the curve of reigns is 0 to 103·1 years, and from the above curve 110·40 years, running from $-1{\cdot}84$ to $+108{\cdot}56$ years, not a bad approximation. The curve is shown in Diagram II, and it looks more nearly normal than it really is, when compared with the normal curve.

We require to know the area embraced by the tail from $257/7 = 36{\cdot}7143$ years onwards. This was obtained by plotting the tail of the curve on a big scale and mechanically integrating it :

$$\text{Area of tail/total area} = {\cdot}00496,$$

or, the probability of seven kings reigning for 257 years is very nearly ·005.

DIAGRAM II DISTRIBUTION OF MEANS IN SAMPLES OF SEVEN REIGNS



This value is badly out of accord with that (·00013) found from the rougher method with data from the German Emperors (see our p. 246 ftn.), but is of the same order with what may be deduced from Condorcet's result when we apply a geometrical series of radix $\frac{1}{2}$, i.e. ·0016 (see our p. 248 ftn.).

If, instead of actually determining the skew curve for the distribution of means, we had, judging from B_1 and B_2, held that it might be considered approximately normal, we should have

$$\frac{\text{Deviation of Mean Reign of Seven Kings}}{\text{Standard Deviation of Sampled Means}} = \frac{36{\cdot}7143 - 21{\cdot}2520}{5{\cdot}39328} = 2{\cdot}87,$$

which gives us, for area of normal curve up to 36·7143, $\frac{1}{2}(1+\alpha) = {\cdot}99795$, or a probability of ·00205 for the length of the seven reigns covering 257 years or more. It will be seen at once from this result that the skewness in the distribution of means of small samples drawn from a skew population can be influential in problems of the present kind.

2*nd Method.* Here the problem to be solved is treated as a purely actuarial one, and it seemed to me best to seek actuarial aid in its solution. My friend Mr W. Palin Elderton, to whom I applied, placed the problem in the hands of Mr Charles F. Trustam, and he has most kindly provided the following discussion and approximate solution. While the actuarial method may thus be usefully applied to historical problems, it appears to me possible that the method of small samples might occasionally be of value to the actuary, a curve of frequency of after-lives replacing that of reigns.

Solution by Charles F. Trustam.

Seven men are chosen at random between the ages of 30 *and* 60. *Find the chance that their total future lifetime will be not less than* 257 *years.*

Some assumption must be made here regarding the age-distribution of the group from which the lives are chosen. It will be assumed in what follows, for the purposes of illustration, that the lives are distributed according to the l_x column of the mortality table which represents the mortality experienced by the group.

On this basis, the chance that a man chosen at random from the group between the ages of 30 and 60 will be aged x is

$$\frac{l_x}{l_{30} + l_{31} + \ldots + l_{59}},$$

and the chance that such a life will be aged x and will survive n years is

$$\frac{l_x}{l_{30} + \ldots + l_{59}} \cdot \frac{l_{x+n}}{l_x} = \frac{l_{x+n}}{l_{30} + \ldots + l_{59}}.$$

Hence the total chance that a life chosen at random between ages 30 and 60 will survive n years is

$$\frac{l_{30+n} + \ldots + l_{59+n}}{l_{30} + \ldots + l_{59}},$$

which may be written

$$\frac{\lambda_n}{\lambda_0}.$$

Hence the problem reduces to the following:

Given 7 lives aged 0, who survive according to the table λ, find the chance that their total future lifetime will be not less than 257 years.

Let P be the required probability.

Then clearly $$P = ({}_{\frac{257}{7}}p_0)^7 + 7\int_0^{\frac{257}{7}} {}_{n_1}p_0{}^7 \cdot \mu_{n_1} \cdot C_6 \cdot dn_1,$$

where p and μ are the usual functions, based on the λ table, and C_6 is the chance that 6 lives, all aged n_1, will have a total future lifetime of not less than $(257 - 7n_1)$ years.

Similarly

$$C_6 = ({}_{\frac{257-7n_1}{6}}p_{n_1})^6 + 6\int_{n_1}^{n_1+\frac{257-7n_1}{6}} {}_{n_2-n_1}p_{n_1}{}^6 \cdot \mu_{n_2} \cdot C_5 \cdot dn_2$$

$$= ({}_{\frac{257-7n_1}{6}}p_{n_1})^6 + 6\int_{n_1}^{\frac{257-n_1}{6}} {}_{n_2-n_1}p_{n_1}{}^6 \cdot \mu_{n_2} \cdot C_5 \cdot dn_2,$$

$$C_5 = ({}_{\frac{257-n_1-6n_2}{5}}p_{n_2})^5 + 5\int_{n_2}^{\frac{257-n_1-n_2}{5}} {}_{n_3-n_2}p_{n_2}{}^5 \cdot \mu_{n_3} \cdot C_4 \cdot dn_3,$$

$$C_4 = ({}_{\frac{257-n_1-n_2-5n_3}{4}}p_{n_3})^4 + 4\int_{n_3}^{\frac{257-n_1-n_2-n_3}{4}} {}_{n_4-n_3}p_{n_3}{}^4 \cdot \mu_{n_4} \cdot C_3 \cdot dn_4,$$

$$C_3 = ({}_{\frac{257-n_1-n_2-n_3-4n_4}{3}}p_{n_4})^3 + 3\int_{n_4}^{\frac{257-n_1-n_2-n_3-n_4}{3}} {}_{n_5-n_4}p_{n_4}{}^3 \cdot \mu_{n_5} \cdot C_2 \cdot dn_5,$$

$$C_2 = ({}_{\frac{257-n_1-n_2-n_3-n_4-3n_5}{2}}p_{n_5})^2 + 2\int_{n_5}^{\frac{257-n_1-n_2-n_3-n_4-n_5}{2}} {}_{n_6-n_5}p_{n_5}{}^2 \cdot \mu_{n_6} \cdot C_1 \cdot dn_6,$$

and $$C_1 = {}_{257-n_1-n_2-n_3-n_4-n_5-2n_6}p_{n_6}.$$

Hence, by successive substitution and integration we may, in theory, find P:

$$C_1 = \frac{\lambda_{257-n_1-n_2-n_3-n_4-n_5-n_6}}{\lambda_{n_6}},$$

$$C_2 = \left[\frac{\lambda_{\frac{257-n_1-n_2-n_3-n_4-n_5}{2}}}{\lambda_{n_5}}\right]^2 + 2\int_{n_5}^{\frac{257-n_1-n_2-n_3-n_4-n_5}{2}} -\frac{\lambda_{n_6}{}^2}{\lambda_{n_5}{}^2}\cdot\frac{1}{\lambda_{n_6}}\cdot d\lambda_{n_6}\cdot\frac{\lambda_{257-n_1-n_2-n_3-n_4-n_5-n_6}}{\lambda_{n_6}}$$

$$= \left[\frac{\lambda_{\frac{257-n_1-n_2-n_3-n_4-n_5}{2}}}{\lambda_{n_5}}\right]^2 + \frac{2}{\lambda_{n_5}{}^2}\int_{n_5}^{\frac{257-n_1-n_2-n_3-n_4-n_5}{2}} -\lambda_{257-n_1-n_2-n_3-n_4-n_5-n_6}\,d\lambda_{n_6},$$

$$C_3 = \left[\frac{\lambda_{\frac{257-n_1-n_2-n_3-n_4}{3}}}{\lambda_{n_4}}\right]^3 + 3\int_{n_4}^{\frac{257-n_1-n_2-n_3-n_4}{3}} \frac{\lambda_{n_5}{}^3}{\lambda_{n_4}{}^3}\cdot-\frac{1}{\lambda_{n_5}}\cdot d\lambda_{n_5}\cdot\frac{\lambda^2{}_{\frac{257-n_1-n_2-n_3-n_4-n_5}{2}}}{\lambda_{n_5}{}^2} + \text{etc.}$$

$$= \left[\frac{\lambda_{\frac{257-n_1-n_2-n_3-n_4}{3}}}{\lambda_{n_4}}\right]^3 + \frac{3}{\lambda_{n_4}{}^3}\int_{n_4}^{\frac{257-n_1-n_2-n_3-n_4}{3}} -\lambda^2{}_{\frac{257-n_1-n_2-n_3-n_4-n_5}{2}}\,d\lambda_{n_5}$$

$$+\frac{3\cdot 2}{\lambda_{n_4}{}^3}\int_{n_4}^{\frac{257-n_1-n_2-n_3-n_4}{3}} -d\lambda_{n_5}\int_{n_5}^{\frac{257-n_1-n_2-n_3-n_4-n_5}{2}} -\lambda_{257-n_1-n_2-n_3-n_4-n_5-n_6}\,d\lambda_{n_6},$$

$$C_4 = \left(\frac{\lambda_{\frac{257-n_1-n_2-n_3}{4}}}{\lambda_{n_3}}\right)^4 + 4\int_{n_3}^{\frac{257-n_1-n_2-n_3}{4}} \frac{\lambda_{n_4}^4}{\lambda_{n_3}^4}\cdot -\frac{1}{\lambda_{n_4}}\,.\,d\lambda_{n_4}\cdot \frac{\lambda^3_{\frac{257-n_1-n_2-n_3-n_4}{3}}}{\lambda_{n_4}^3}$$

$$+ 4\int_{n_3}^{\frac{257-n_1-n_2-n_3}{4}} \frac{\lambda_{n_4}^4}{\lambda_{n_3}^4}\cdot -\frac{1}{\lambda_{n_4}}\,.\,d\lambda_{n_4}\cdot \frac{3}{\lambda_{n_4}^3}\cdot \int_{n_4}^{\frac{257-n_1-n_2-n_3-n_4}{3}} - \lambda^2_{\frac{257-n_1-n_2-n_3-n_4-n_5}{2}}\,d\lambda_{n_5}$$

$$+ 4\int_{n_3}^{\frac{257-n_1-n_2-n_3}{4}} \frac{\lambda_{n_4}^4}{\lambda_{n_3}^4}\cdot -\frac{1}{\lambda_{n_4}}\,.\,d\lambda_{n_4}\cdot \frac{3\,.\,2}{\lambda_{n_4}^3}\int_{n_4}^{\frac{257-n_1-n_2-n_3-n_4}{3}} - d\lambda_{n_5}$$

$$\times \int_{n_5}^{\frac{257-n_1-n_2-n_3-n_4-n_5}{2}} - \lambda_{257-n_1-n_2-n_3-n_4-n_5-n_6}\,d\lambda_{n_6}$$

$$= \left[\frac{\lambda_{\frac{257-n_1-n_2-n_3}{4}}}{\lambda_{n_3}}\right]^4 + \frac{4}{\lambda_{n_3}^4}\int_{n_3}^{\frac{257-n_1-n_2-n_3}{4}} - \lambda^3_{\frac{257-n_1-n_2-n_3-n_4}{3}}\,d\lambda_{n_4}$$

$$+ \frac{4\,.\,3}{\lambda_{n_3}^4}\int_{n_3}^{\frac{257-n_1-n_2-n_3}{4}} - d\lambda_{n_4}\int_{n_4}^{\frac{257-n_1-n_2-n_3-n_4}{3}} - \lambda^2_{\frac{257-n_1-n_2-n_3-n_4-n_5}{2}}\,d\lambda_{n_5}$$

$$+ \frac{4\,.\,3\,.\,2}{\lambda_{n_3}^4}\int_{n_3}^{\frac{257-n_1-n_2-n_3}{4}} - d\lambda_{n_4}\int_{n_4}^{\frac{257-n_1-n_2-n_3-n_4}{3}} - d\lambda_{n_5}$$

$$\times \int_{n_5}^{\frac{257-n_1-n_2-n_3-n_4-n_5}{2}} - \lambda_{257-n_1-n_2-n_3-n_4-n_5-n_6}\,d\lambda_{n_6},$$

$$C_5 = \left[\frac{\lambda_{\frac{257-n_1-n_2}{5}}}{\lambda_{n_2}}\right]^5 + \frac{5}{\lambda_{n_2}^5}\int_{n_2}^{\frac{257-n_1-n_2}{5}} \lambda_{n_3}^4\,.\,C_4\,.\,-\,d\lambda_{n_3}$$

$$= \left[\frac{\lambda_{\frac{257-n_1-n_2}{5}}}{\lambda_{n_2}}\right]^5 + \frac{5}{\lambda_{n_2}^5}\int_{n_2}^{\frac{257-n_1-n_2}{5}} - \lambda^4_{\frac{257-n_1-n_2-n_3}{4}}\,d\lambda_{n_3}$$

$$+ \frac{5\,.\,4}{\lambda_{n_2}^5}\int_{n_2}^{\frac{257-n_1-n_2}{5}} - d\lambda_{n_3}\int_{n_3}^{\frac{257-n_1-n_2-n_3}{4}} - \lambda^3_{\frac{257-n_1-n_2-n_3-n_4}{3}}\,d\lambda_{n_4}$$

$$+ \frac{5\,.\,4\,.\,3}{\lambda_{n_2}^5}\int_{n_2}^{\frac{257-n_1-n_2}{5}} - d\lambda_{n_3}\int_{n_3}^{\frac{257-n_1-n_2-n_3}{4}} - d\lambda_{n_4}$$

$$\times \int_{n_4}^{\frac{257-n_1-n_2-n_3-n_4}{3}} - \lambda^2_{\frac{257-n_1-n_2-n_3-n_4-n_5}{2}}\,d\lambda_{n_5}$$

$$+ \frac{5\,.\,4\,.\,3\,.\,2}{\lambda_{n_2}^5}\int_{n_2}^{\frac{257-n_1-n_2}{5}} - d\lambda_{n_3}\int_{n_3}^{\frac{257-n_1-n_2-n_3}{4}} - d\lambda_{n_4}\int_{n_4}^{\frac{257-n_1-n_2-n_3-n_4}{3}} - d\lambda_{n_5}$$

$$\times \int_{n_5}^{\frac{257-n_1-n_2-n_3-n_4-n_5}{2}} - \lambda_{257-n_1-n_2-n_3-n_4-n_5-n_6}\,d\lambda_{n_6},$$

$$C_6 = \left[\frac{\lambda_{\frac{257-n_1}{6}}}{\lambda_{n_1}}\right]^6 + \frac{6}{\lambda_{n_1}{}^6}\int_{n_1}^{\frac{257-n_1}{6}} \lambda_{n_2}{}^5 . C_5 . - d\lambda_{n_2}$$

$$= \left[\frac{\lambda_{\frac{257-n_1}{6}}}{\lambda_{n_1}}\right]^6 + \frac{6}{\lambda_{n_1}{}^6}\int_{n_1}^{\frac{257-n_1}{6}} -\lambda^5{}_{\frac{257-n_1-n_2}{5}} d\lambda_{n_2}$$

$$+ \frac{6.5}{\lambda_{n_1}{}^6}\int_{n_1}^{\frac{257-n_1}{6}} - d\lambda_{n_2}\int_{n_2}^{\frac{257-n_1-n_2}{5}} - \lambda^4{}_{\frac{257-n_1-n_2-n_3}{4}} d\lambda_{n_3}$$

$$+ \frac{6.5.4}{\lambda_{n_1}{}^6}\int_{n_1}^{\frac{257-n_1}{6}} - d\lambda_{n_2}\int_{n_2}^{\frac{257-n_1-n_2}{5}} - d\lambda_{n_3}\int_{n_3}^{\frac{257-n_1-n_2-n_3}{4}} - \lambda^3{}_{\frac{257-n_1-n_2-n_3-n_4}{3}} d\lambda_{n_4}$$

$$+ \frac{6.5.4.3}{\lambda_{n_1}{}^6}\int_{n_1}^{\frac{257-n_1}{6}} - d\lambda_{n_2}\int_{n_2}^{\frac{257-n_1-n_2}{5}} - d\lambda_{n_3}\int_{n_3}^{\frac{257-n_1-n_2-n_3}{4}} - d\lambda_{n_4}$$

$$\times \int_{n_4}^{\frac{257-n_1-n_2-n_3-n_4}{3}} - \lambda^2{}_{\frac{257-n_1-n_2-n_3-n_4-n_5}{2}} d\lambda_{n_5}$$

$$+ \frac{6.5.4.3.2}{\lambda_{n_1}{}^6}\int_{n_1}^{\frac{257-n_1}{6}} - d\lambda_{n_2}\int_{n_2}^{\frac{257-n_1-n_2}{5}} - d\lambda_{n_3}\int_{n_3}^{\frac{257-n_1-n_2-n_3}{4}} - d\lambda_{n_4}$$

$$\times \int_{n_4}^{\frac{257-n_1-n_2-n_3-n_4}{3}} - d\lambda_{n_5}\int_{n_5}^{\frac{257-n_1-n_2-n_3-n_4-n_5}{2}} - \lambda_{257-n_1-n_2-n_3-n_4-n_5-n_6} d\lambda_{n_6}.$$

Hence $P = \left[\frac{\lambda_{\frac{257}{7}}}{\lambda_0}\right]^7 + \frac{7}{\lambda_0{}^7}\int_0^{\frac{257}{7}} \lambda_{n_1}{}^6 . C_6 . - d\lambda_{n_1}$

$$= \left[\frac{\lambda_{\frac{257}{7}}}{\lambda_0}\right]^7 + \frac{7}{\lambda_0{}^7}\int_0^{\frac{257}{7}} - \lambda^6{}_{\frac{257-n_1}{6}} d\lambda_{n_1} + \frac{7.6}{\lambda_0{}^7}\int_0^{\frac{257}{7}} - d\lambda_{n_1}\int_{n_1}^{\frac{257-n_1}{6}} - \lambda^5{}_{\frac{257-n_1-n_2}{5}} d\lambda_{n_2}$$

$$+ \frac{7.6.5}{\lambda_0{}^7}\int_0^{\frac{257}{7}} - d\lambda_{n_1}\int_{n_1}^{\frac{257-n_1}{6}} - d\lambda_{n_2}\int_{n_2}^{\frac{257-n_1-n_2}{5}} - \lambda^4{}_{\frac{257-n_1-n_2-n_3}{4}} d\lambda_{n_3}$$

$$+ \frac{7.6.5.4}{\lambda_0{}^7}\int_0^{\frac{257}{7}} - d\lambda_{n_1}\int_{n_1}^{\frac{257-n_1}{6}} - d\lambda_{n_2}\int_{n_2}^{\frac{257-n_1-n_2}{5}} - d\lambda_{n_3}$$

$$\times \int_{n_3}^{\frac{257-n_1-n_2-n_3}{4}} - \lambda^3{}_{\frac{257-n_1-n_2-n_3-n_4}{3}} d\lambda_{n_4}$$

$$+ \frac{7.6.5.4.3}{\lambda_0{}^7}\int_0^{\frac{257}{7}} - d\lambda_{n_1}\int_{n_1}^{\frac{257-n_1}{6}} - d\lambda_{n_2}\int_{n_2}^{\frac{257-n_1-n_2}{5}} - d\lambda_{n_2}$$

$$\times \int_{n_3}^{\frac{257-n_1-n_2-n_3}{4}} - d\lambda_{n_1}\int_{n_4}^{\frac{257-n_1-n_2-n_3-n_4}{3}} - \lambda^2{}_{\frac{257-n_1-n_2-n_3-n_4-n_5}{2}} d\lambda_{n_5}$$

$$+\frac{7!}{\lambda_0{}^7}\int_0^{\frac{257}{7}} -d\lambda_{n_1}\int_{n_1}^{\frac{257-n_1}{6}} -d\lambda_{n_2}\int_{n_2}^{\frac{257-n_1-n_2}{5}} -d\lambda_{n_3}\int_{n_3}^{\frac{257-n_1-n_2-n_3}{4}} -d\lambda_{n_4}$$

$$\times\int_{n_4}^{\frac{257-n_1-n_2-n_3-n_4}{3}} -d\lambda_{n_5}\int_{n_5}^{\frac{257-n_1-n_2-n_3-n_4-n_5}{2}} -\lambda_{257-n_1-n_2-n_3-n_4-n_5-n_6}\,d\lambda_{n_6}$$

$$=\left[\frac{\lambda_{\frac{257}{7}}}{\lambda_0}\right]^7+S_1+S_2+S_3+S_4+S_5+S_6, \text{ say.}$$

$$\lambda_t=\sum_{n=0}^{99} l_{30+t+n}$$

λ_0	2 317 271
λ_5	2 151 752
λ_{10}	1 953 150
λ_{15}	1 715 501
λ_{20}	1 439 785
λ_{25}	1 139 027
λ_{30}	838 330
λ_{35}	565 732
λ_{40}	341 476
λ_{45}	176 212
λ_{50}	72 215
λ_{55}	20 935
λ_{60}	3 577
λ_{65}	270
λ_{70}	5

$$\lambda_{\frac{257}{7}}=\lambda_{36\frac{5}{7}}=482\,906$$

$$\lambda_{\frac{257}{6}}=\lambda_{42\frac{5}{6}}=240\,398$$

$$\left[\frac{\lambda_{\frac{257}{7}}}{\lambda_0}\right]^7=\cdot00001707$$

$$\left[\frac{\lambda_{\frac{257}{6}}}{\lambda_0}\right]^7=\cdot00000013$$

Approximate Solution.

An approximate solution to the foregoing may be obtained as follows:

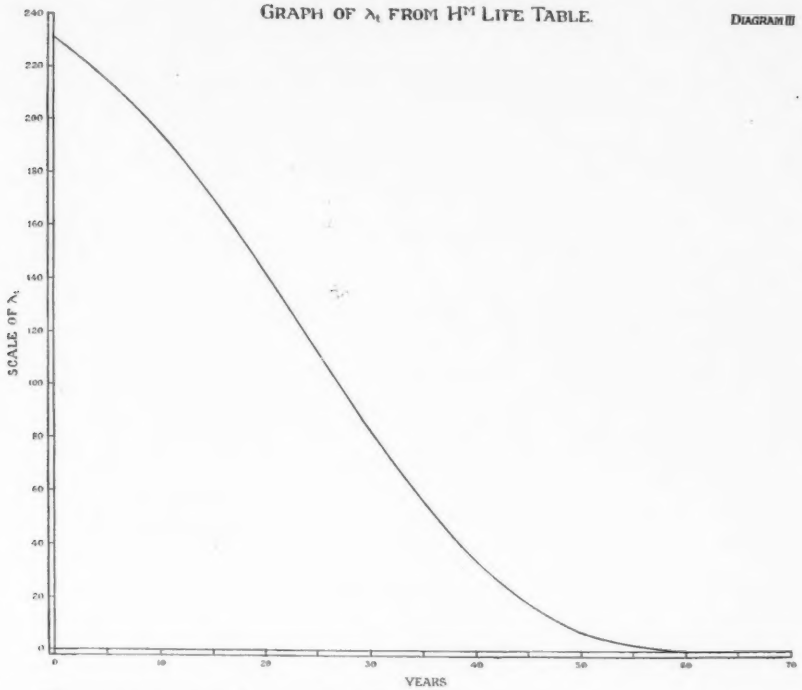
In the first place consider the curve representing the function λ. The illustration given is based on H^M mortality, and it will be seen that as a broad approximation we may assume that for values of λ between 0 and, say, 45 the curve is roughly a straight line.

If now $\lambda_n = an + b$

and $\lambda_{\kappa-\theta n} = a(\kappa - \theta n) + b \qquad (\kappa = \text{const.}),$

then $d\lambda_{\kappa-\theta n} = -\theta a d\lambda = -\theta d\lambda_n.$

GRAPH OF λ_t FROM H^M LIFE TABLE. DIAGRAM III



We may now proceed to evaluate $S_1 \ldots S_6$:

$$S_1 = \frac{7}{\lambda_0{}^7}\int_0^{\frac{257}{7}} \lambda^6_{\frac{257-n_1}{6}} \cdot -d\lambda_{n_1} = \frac{7}{\lambda_0{}^7}\int_{n_1=0}^{\frac{257}{7}} \lambda^6_{\frac{257-n_1}{6}} \cdot 6 \cdot d\lambda_{\frac{257-n_1}{6}}$$

$$= \frac{6}{\lambda_0{}^7}\left[\lambda^7_{\frac{257}{7}} - \lambda^7_{\frac{257}{6}}\right],$$

$$S_2 = \frac{7 \cdot 6}{\lambda_0{}^7}\int_0^{\frac{257}{7}} -d\lambda_{n_1}\int_{n_1}^{\frac{257-n_1}{6}} \lambda^5_{\frac{257-n_1-n_2}{5}} \cdot -d\lambda_{n_2} = \frac{7 \cdot 6}{\lambda_0{}^7}\int_0^{\frac{257}{7}} -d\lambda_{n_1}$$

$$\times \int_{n_2=n_1}^{\frac{257-n_1}{6}} \lambda^5_{\frac{257-n_1-n_2}{5}} \cdot 5 \cdot d\lambda_{\frac{257-n_1-n_2}{5}}$$

$$= \frac{7 \cdot 5}{\lambda_0{}^7}\int_0^{\frac{257}{7}} -d\lambda_{n_1} \cdot \left[\lambda^6_{\frac{257-n_1}{6}} - \lambda^6_{\frac{257-2n_1}{5}}\right]$$

$$= 5S_1 + \frac{7 \cdot 5}{\lambda_0{}^7}\int_0^{\frac{257}{7}} \lambda^6_{\frac{257-2n_1}{5}} \cdot d\lambda_{n_1} = 5S_1 - \frac{5}{\lambda_0{}^7} \cdot \frac{5}{2}\left[\lambda^7_{\frac{257}{7}} - \lambda^7_{\frac{257}{5}}\right],$$

$$S_3 = \frac{7.6.5}{\lambda_0{}^7}\int_0^{\frac{257}{7}} - d\lambda_{n_1}\int_{n_1}^{\frac{257-n_1}{6}} - d\lambda_{n_2}\int_{n_2}^{\frac{257-n_1-n_2}{5}} \lambda^4{}_{\frac{257-n_1-n_2-n_3}{4}} . - d\lambda_{n_3}$$

$$= \frac{7.6.5}{\lambda_0{}^7}\int_0^{\frac{257}{7}} - d\lambda_{n_1}\int_{n_1}^{\frac{257-n_1}{6}} - d\lambda_{n_2}\int_{n_3=n_2}^{\frac{257-n_1-n_2}{5}} \lambda^4{}_{\frac{257-n_1-n_2-n_3}{4}} . 4 . d\lambda_{\frac{257-n_1-n_2-n_3}{4}}$$

$$= \frac{7.6.4}{\lambda_0{}^7}\int_0^{\frac{257}{7}} - d\lambda_{n_1}\int_{n_1}^{\frac{257-n_1}{6}} - d\lambda_{n_2} . \left[\lambda^5{}_{\frac{257-n_1-n_2}{5}} - \lambda^5{}_{\frac{257-n_1-2n_2}{4}}\right]$$

$$= 4S_2 + \frac{7.6.4}{\lambda_0{}^7}\int_0^{\frac{257}{7}} - d\lambda_{n_1}\int_{n_2=n_1}^{\frac{257-n_1}{6}} \lambda^5{}_{\frac{257-n_1-2n_2}{4}} . \frac{4}{2} . - d\lambda_{\frac{257-n_1-2n_2}{4}}$$

$$= 4S_2 - \frac{7.4}{\lambda_0{}^7}\cdot\frac{4}{2}\int_0^{\frac{257}{7}} - d\lambda_{n_1} . \left[\lambda^6{}_{\frac{257-n_1}{6}} - \lambda^6{}_{\frac{257-3n_1}{4}}\right]$$

$$= 4S_2 - \frac{4^2}{2} . S_1 + \frac{7.4}{\lambda_0{}^7}\cdot\frac{4}{2}\int_0^{\frac{257}{7}} \lambda^6{}_{\frac{257-3n_1}{4}} . \frac{4}{3} . d\lambda_{\frac{257-3n_1}{4}}$$

$$= 4S_2 - \frac{4^2}{2} . S_1 + \frac{4^3}{3.2}\cdot\frac{1}{\lambda_0{}^7}\left[\lambda^7{}_{\frac{257}{7}} - \lambda^7{}_{\frac{257}{4}}\right],$$

$$S_4 = \frac{7.6.5.4}{\lambda_0{}^7}\int_0^{\frac{257}{7}} - d\lambda_{n_1}\int_{n_1}^{\frac{257-n_1}{6}} - d\lambda_{n_2}\int_{n_2}^{\frac{257-n_1-n_2}{5}} - d\lambda_{n_3}$$
$$\times\int_{n_3}^{\frac{257-n_1-n_2-n_3}{4}} \lambda^3{}_{\frac{257-n_1-n_2-n_3-n_4}{3}} . - d\lambda_{n_4}$$

$$= \frac{7.6.5.4}{\lambda_0{}^7}\int_0^{\frac{257}{7}} - d\lambda_{n_1}\int_{n_1}^{\frac{257-n_1}{6}} - d\lambda_{n_2}\int_{n_2}^{\frac{257-n_1-n_2}{5}} - d\lambda_{n_3}$$
$$\times\int_{n_3}^{\frac{257-n_1-n_2-n_3}{4}} \lambda^3{}_{\frac{257-n_1-n_2-n_3-n_4}{3}} . 3 . d\lambda_{\frac{257-n_1-n_2-n_3-n_4}{3}}$$

$$= \frac{7.6.5.3}{\lambda_0{}^7}\int_0^{\frac{257}{7}} - d\lambda_{n_1}\int_{n_1}^{\frac{257-n_1}{6}} - d\lambda_{n_2}\int_{n_2}^{\frac{257-n_1-n_2}{5}} - d\lambda_{n_3}$$
$$\times\left[\lambda^4{}_{\frac{257-n_1-n_2-n_3}{4}} - \lambda^4{}_{\frac{257-n_1-n_2-2n_3}{3}}\right]$$

$$= 3S_3 - \frac{7.6.5.3}{\lambda_0{}^7}\int_0^{\frac{257}{7}} - d\lambda_{n_1}\int_{n_1}^{\frac{257-n_1}{6}} - d\lambda_{n_2}$$
$$\times\int_{n_2}^{\frac{257-n_1-n_2}{5}} \lambda^4{}_{\frac{257-n_1-n_2-2n_3}{3}} . - d\lambda_{n_3}$$

$$= 3S_3 - \frac{7.6.5.3}{\lambda_0{}^7}\int_0^{\frac{257}{7}} - d\lambda_{n_1}\int_{n_1}^{\frac{257-n_1}{6}} - d\lambda_{n_2}$$
$$\times\int_{n_2}^{\frac{257-n_1-n_2}{5}} \lambda^4{}_{\frac{257-n_1-n_2-2n_3}{5}} . \frac{3}{2} . d\lambda_{\frac{257-n_1-n_2-2n_3}{3}}$$

$$= 3S_3 - \frac{7.6}{{\lambda_0}^7} \cdot \frac{3^2}{2} \int_0^{\frac{257}{7}} - d\lambda_{n_1} \int_{n_1}^{\frac{257-n_1}{6}} - d\lambda_{n_2} \cdot \left[{\lambda^5}_{\frac{257-n_1-n_2}{3}} - {\lambda^5}_{\frac{257-n_1-3n_2}{3}} \right]$$

$$= 3S_3 - \frac{3^2}{2} \cdot S_2 + \frac{7.6}{{\lambda_0}^7} \cdot \frac{3^2}{2} \int_0^{\frac{257}{7}} - d\lambda_{n_1} \int_{n_1}^{\frac{257-n_1}{6}} {\lambda^5}_{\frac{257-n_1-3n_2}{3}} \cdot 1 \cdot d\lambda_{\frac{257-n_1-3n_2}{3}}$$

$$= 3S_3 - \frac{3^2}{2} \cdot S_2 + \frac{7}{{\lambda_0}^7} \cdot \frac{3^2}{2} \int_0^{\frac{257}{7}} - d\lambda_{n_1} \cdot \left[{\lambda^6}_{\frac{257-n_1}{6}} - {\lambda^6}_{\frac{257-4n_1}{3}} \right]$$

$$= 3S_3 - \frac{3^2}{2} \cdot S_2 + \frac{3^2}{2} \cdot S_1 - \frac{7}{{\lambda_0}^7} \cdot \frac{3^2}{2} \cdot \int_0^{\frac{257}{7}} {\lambda^6}_{\frac{257-4n_1}{3}} \cdot \frac{3}{4} \cdot d\lambda_{\frac{257-4n_1}{3}}$$

$$= 3S_3 - \frac{3^2}{2} \cdot S_2 + \frac{3^2}{2} \cdot S_1 - \frac{3^3}{8{\lambda_0}^7} \left[{\lambda^7}_{\frac{257}{7}} - {\lambda^7}_{\frac{257}{3}} \right]$$

$$= 3S_3 - \frac{3^2}{2} \cdot S_2 + \frac{3^3}{2.3} \cdot S_1 - \frac{3^4}{2.3.4.{\lambda_0}^7} \left[{\lambda^7}_{\frac{257}{7}} - {\lambda^7}_{\frac{257}{3}} \right],$$

$$S_5 = \frac{7.6.5.4.3}{{\lambda_0}^7} \int_0^{\frac{257}{7}} - d\lambda_{n_1} \int_{n_1}^{\frac{257-n_1}{6}} - d\lambda_{n_2} \int_{n_2}^{\frac{257-n_1-n_2}{5}} - d\lambda_{n_3} \int_{n_3}^{\frac{257-n_1-n_2-n_3}{4}} - d\lambda_{n_4}$$
$$\times \int_{n_4}^{\frac{257-n_1-n_2-n_3-n_4}{3}} {\lambda^2}_{\frac{257-n_1-n_2-n_3-n_4-n_5}{2}} \cdot 2 \cdot d\lambda_{\frac{257-n_1-n_2-n_3-n_4-n_5}{2}}$$

$$= \frac{7.6.5.4.2}{{\lambda_0}^7} \int_0^{\frac{257}{7}} - d\lambda_{n_1} \int_{n_1}^{\frac{257-n_1}{6}} - d\lambda_{n_2} \int_{n_2}^{\frac{257-n_1-n_2}{5}} - d\lambda_{n_3} \int_{n_3}^{\frac{257-n_1-n_2-n_3}{4}} - d\lambda_{n_4}$$
$$\times \left[{\lambda^3}_{\frac{257-n_1-n_2-n_3-n_4}{3}} - {\lambda^3}_{\frac{257-n_1-n_2-n_3-2n_4}{2}} \right]$$

$$= 2S_4 - \frac{7.6.5.4.2}{{\lambda_0}^7} \int_0^{\frac{257}{7}} - d\lambda_{n_1} \int_{n_1}^{\frac{257-n_1}{6}} - d\lambda_{n_2} \int_{n_2}^{\frac{257-n_1-n_2}{5}} - d\lambda_{n_3}$$
$$\times \int_{n_3}^{\frac{257-n_1-n_2-n_3}{4}} {\lambda^3}_{\frac{257-n_1-n_2-n_3-2n_4}{2}} \cdot 1 \cdot d\lambda_{\frac{257-n_1-n_2-n_3-2n_4}{2}}$$

$$= 2S_4 - \frac{7.6.5.2}{{\lambda_0}^7} \int_0^{\frac{257}{7}} - d\lambda_{n_1} \int_{n_1}^{\frac{257-n_1}{6}} - d\lambda_{n_2} \int_{n_2}^{\frac{257-n_1-n_2}{5}} - d\lambda_{n_3}$$
$$\times \left[{\lambda^4}_{\frac{257-n_1-n_2-n_3}{4}} - {\lambda^4}_{\frac{257-n_1-n_3-3n_2}{2}} \right]$$

$$= 2S_4 - 2S_3 + \frac{7.6.5.2}{{\lambda_0}^7} \int_0^{\frac{257}{7}} - d\lambda_{n_1} \int_{n_1}^{\frac{257-n_1}{6}} - d\lambda_{n_2}$$
$$\times \int_{n_2}^{\frac{257-n_1-n_2}{5}} {\lambda^4}_{\frac{257-n_1-n_2-3n_2}{2}} \cdot \frac{2}{3} \cdot d\lambda_{\frac{257-n_1-n_2-3n_2}{2}}$$

$$= 2S_4 - 2S_3 + \frac{7.6}{{\lambda_0}^7} \cdot \frac{2^2}{3} \int_0^{\frac{257}{7}} - d\lambda_{n_1} \int_{n_1}^{\frac{257-n_1}{6}} - d\lambda_{n_2} \left[{\lambda^5}_{\frac{257-n_1-n_2}{5}} - {\lambda^5}_{\frac{257-n_1-4n_2}{2}} \right]$$

$$= 2S_4 - 2S_3 + \frac{2^2}{3} \cdot S_2 - \frac{7 \cdot 6}{{\lambda_0}^7} \cdot \frac{2^2}{3} \int_0^{\frac{257}{7}} - d\lambda_{n_1} \int_{n_1}^{\frac{257-n_1}{6}} \lambda^5{}_{\frac{257-n_1-4n_2}{2}} \cdot \tfrac{2}{4} \cdot d\lambda_{\frac{257-n_1-4n_2}{2}}$$

$$= 2S_4 - 2S_3 + \frac{2^2}{3} \cdot S_2 - \frac{7}{{\lambda_0}^7} \cdot \frac{2^3}{3 \cdot 4} \int_0^{\frac{257}{7}} - d\lambda_{n_1} \left[\lambda^6{}_{\frac{257-n_1}{6}} - \lambda^6{}_{\frac{257-5n_1}{2}} \right]$$

$$= 2S_4 - 2S_3 + \frac{2^2}{3} \cdot S_2 - \frac{2^3}{3 \cdot 4} \cdot S_1 + \frac{7}{{\lambda_0}^7} \cdot \frac{2^3}{3 \cdot 4} \int_0^{\frac{257}{7}} \lambda^6{}_{\frac{257-5n_1}{2}} \cdot \tfrac{2}{5} \cdot d\lambda_{\frac{257-5n_1}{2}}$$

$$= 2S_4 - 2S_3 + \frac{2^2}{3} \cdot S_2 - \frac{2^3}{3 \cdot 4} \cdot S_1 + \frac{2^4}{3 \cdot 4 \cdot 5 \cdot {\lambda_0}^7} \left[\lambda^7{}_{\frac{257}{7}} - \lambda^7{}_{\frac{257}{2}} \right]$$

$$= 2S_4 - \frac{2^2}{2} \cdot S_3 + \frac{2^3}{2 \cdot 3} \cdot S_2 - \frac{2^4}{2 \cdot 3 \cdot 4} \cdot S_1 + \frac{2^5}{2 \cdot 3 \cdot 4 \cdot 5 \cdot {\lambda_0}^7} \left[\lambda^7{}_{\frac{257}{7}} - \lambda^7{}_{\frac{257}{2}} \right],$$

$$S_6 = \frac{7!}{{\lambda_0}^7} \int_0^{\frac{257}{7}} - d\lambda_{n_1} \int_{n_1}^{\frac{257-n_1}{6}} - d\lambda_{n_2} \int_{n_2}^{\frac{257-n_1-n_2}{5}} - d\lambda_{n_3} \int_{n_3}^{\frac{257-n_1-n_2-n_3}{4}} - d\lambda_{n_4}$$

$$\times \int_{n_4}^{\frac{257-n_1-n_2-n_3-n_4}{3}} - d\lambda_{n_5} \int_{n_5}^{\frac{257-n_1-n_2-n_3-n_4-n_5}{2}} \lambda_{257-n_1-n_2-n_3-n_4-n_5} d\lambda_{257-\ldots-n_6}$$

$$= \frac{7 \cdot 6 \cdot 5 \cdot 4 \cdot 3}{{\lambda_0}^7} \int_0^{\frac{257}{7}} - d\lambda_{n_1} \int_{n_1}^{\frac{257-n_1}{6}} - d\lambda_{n_2} \int_{n_2}^{\frac{257-n_1-n_2}{5}} - d\lambda_{n_3} \int_{n_3}^{\frac{257-n_1-n_2-n_3}{4}} - d\lambda_{n_4}$$

$$\times \int_{n_4}^{\frac{257-n_1-n_2-n_3-n_4}{3}} - d\lambda_{n_5} \left[\lambda^2{}_{\frac{257-n_1-n_2-n_3-n_4-n_5}{2}} - \lambda^2{}_{257-n_1-n_2-n_3-n_4-2n_5} \right]$$

$$= S_5 - \frac{7 \cdot 6 \cdot 5 \cdot 4 \cdot 3}{{\lambda_0}^7} \int_0^{\frac{257}{7}} - d\lambda_{n_1} \int_{n_1}^{\frac{257-n_1}{6}} - d\lambda_{n_2} \int_{n_2}^{\frac{257-n_1-n_2}{5}} - d\lambda_{n_3} \int_{n_3}^{\frac{257-n_1-n_2-n_3}{4}} - d\lambda_{n_4}$$

$$\times \int_{n_4}^{\frac{257-n_1-n_2-n_3-n_4}{3}} \lambda^2{}_{257-n_1-n_2-n_3-n_4-2n_5} \cdot \tfrac{1}{2} \cdot d\lambda_{257-n_1-n_2-n_3-n_4-2n_5}$$

$$= S_5 - \frac{7 \cdot 6 \cdot 5 \cdot 4}{2{\lambda_0}^7} \int_0^{\frac{257}{7}} - d\lambda_{n_1} \int_{n_1}^{\frac{257-n_1}{6}} - d\lambda_{n_2} \int_{n_2}^{\frac{257-n_1-n_2}{5}} - d\lambda_{n_3} \int_{n_3}^{\frac{257-n_1-n_2-n_3}{4}} - d\lambda_{n_4}$$

$$\times \left[\lambda^3{}_{\frac{257-n_1-n_2-n_3-n_4}{3}} - \lambda^3{}_{257-n_1-n_2-n_3-3n_4} \right]$$

$$= S_5 - \tfrac{1}{2} S_4 + \frac{7 \cdot 6 \cdot 5 \cdot 4}{2{\lambda_0}^7} \int_0^{\frac{257}{7}} - d\lambda_{n_1} \int_{n_1}^{\frac{257-n_1}{6}} - d\lambda_{n_2} \int_{n_2}^{\frac{257-n_1-n_2}{5}} - d\lambda_{n_3}$$

$$\times \int_{n_3}^{\frac{257-n_1-n_2-n_3}{4}} \lambda^3{}_{257-n_1-n_2-n_3-3n_4} \cdot \tfrac{1}{3} \cdot d\lambda_{257-n_1-n_2-n_3-3n_4}$$

$$= S_5 - \tfrac{1}{2} S_4 + \frac{7 \cdot 6 \cdot 5}{3 \cdot 2{\lambda_0}^7} \int_0^{\frac{257}{7}} - d\lambda_{n_1} \int_{n_1}^{\frac{257-n_1}{6}} - d\lambda_{n_2} \int_{n_2}^{\frac{257-n_1-n_2}{5}} - d\lambda_{n_3}$$

$$\times \left[\lambda^4{}_{\frac{257-n_1-n_2-n_3}{4}} - \lambda^4{}_{257-n_1-n_2-4n_3} \right]$$

$$= S_5 - \frac{S_4}{2} + \frac{S_3}{2.3} - \frac{7.6.5}{3.2.\lambda_0^7}\int_0^{\frac{257}{7}} - d\lambda_{n_1}\int_{n_1}^{\frac{257-n_1}{6}} - d\lambda_{n_2}$$

$$\times \int_{n_2}^{\frac{257-n_1-n_2}{5}} \lambda^4_{257-n_1-n_2-4n_3} \cdot d\lambda_{257-n_1-n_2-4n_3} \cdot \tfrac{1}{4}$$

$$= S_5 - \frac{S_4}{2} + \frac{S_3}{2.3} - \frac{7.6}{4.3.2.\lambda_0^7}\int_0^{\frac{257}{7}} - d\lambda_{n_1}\int_{n_1}^{\frac{257-n_1}{6}} - d\lambda_{n_2}\left[\lambda^5_{\frac{257-n_1-n_2}{5}} - \lambda^5_{257-n_1-5n_2}\right]$$

$$= S_5 - \frac{S_4}{2} + \frac{S_3}{2.3} - \frac{S_2}{2.3.4} + \frac{7.6}{4.3.2.\lambda_0^7}\int_0^{\frac{257}{7}} - d\lambda_{n_1}$$

$$\times \int_{n_1}^{\frac{257-n_1}{6}} \lambda^5_{257-n_1-5n_2} \cdot \tfrac{1}{5} \cdot d\lambda_{257-n_1-5n_2}$$

$$= S_5 - \frac{S_4}{2} + \frac{S_3}{2.3} - \frac{S_2}{2.3.4} + \frac{7}{5.4.3.2.\lambda_0^7}\int_0^{\frac{257}{7}} - d\lambda_{n_1}\left[\lambda^6_{\frac{257-n_1}{6}} - \lambda^6_{257-6n_1}\right]$$

$$= S_5 - \frac{S_4}{2} + \frac{S_3}{2.3} - \frac{S_2}{2.3.4} + \frac{S_1}{2.3.4.5} - \frac{7}{5.4.3.2.\lambda_0^7}\int_0^{\frac{257}{7}} \lambda^6_{257-6n_1} \cdot \tfrac{1}{6} \cdot d\lambda_{257-6n_1}$$

$$= S_5 - \frac{S_4}{2} + \frac{S_3}{2.3} - \frac{S_2}{2.3.4} + \frac{S_1}{2.3.4.5} - \frac{1}{2.3.4.5.6.\lambda_0^7}\left[\lambda^7_{\frac{257}{7}} - \lambda^7_{257}\right].$$

Numerical Solution.

We have $P = \left[\frac{\lambda_{\frac{257}{7}}}{\lambda_0}\right]^7 + S_1 + S_2 + S_3 + S_4 + S_5 + S_6.$

$$\left[\frac{\lambda_{\frac{257}{7}}}{\lambda_0}\right]^7 = \underline{\cdot 00001707}.$$

$$S_1 = \frac{6}{\lambda_0^7}\left[\lambda^7_{\frac{257}{7}} - \lambda^7_{\frac{257}{6}}\right] = 6\,(\cdot 00001694) = \underline{\cdot 00010164}.$$

$$S_2 = 5S_1 - \tfrac{25}{2}\left[\frac{\lambda_{\frac{257}{7}}}{\lambda_0}\right]^7 \text{ approx.} = \cdot 00050820 - \cdot 00021338 = \underline{\cdot 00029482}.$$

$$S_3 = 4S_2 - 8S_1 + \tfrac{32}{3}\left[\frac{\lambda_{\frac{257}{7}}}{\lambda_0}\right]^7$$

$$= \cdot 00117928 - \cdot 00081312 + \cdot 00018208 = \underline{\cdot 00054824}.$$

$$S_4 = 3S_3 - \tfrac{8}{2}S_2 + \tfrac{9}{2}S_1 - \tfrac{27}{8}\left[\frac{\lambda_{\frac{257}{7}}}{\lambda_0}\right]^7$$

$$= \cdot 00164472 - \cdot 00132669 + \cdot 00045738 - \cdot 00005761 = \underline{\cdot 00071780}.$$

$$S_5 = 2S_4 - 2S_3 + \tfrac{4}{3}S_2 - \tfrac{8}{3}S_1 + \tfrac{4}{15}\left[\frac{\lambda_{257}}{\frac{7}{\lambda_0}}\right]^7$$

$$= \cdot00143560 - \cdot00109648 + \cdot00039309$$

$$- \cdot00006776 + \cdot00000455 = \underline{\cdot00066900}.$$

$$S_6 = S_5 - \tfrac{1}{2}S_4 + \tfrac{1}{6}S_3 - \tfrac{1}{24}S_2 + \tfrac{1}{120}S_1 - \tfrac{1}{720}\left[\frac{\lambda_{257}}{\frac{7}{\lambda_0}}\right]^7$$

$$= \cdot00066900 - \cdot00035890 + \cdot00009137 - \cdot00001228$$

$$+ \cdot00000085 - \cdot00000002 = \underline{\cdot00039002}.$$

Hence $\underline{P = \cdot00274}$. Thus far Mr Trustam.

The three methods we have used for determining the probability of the seven Kings of Rome having reigned for 257 years give values of ·002, ·003 and ·005, all far smaller than Condorcet's ·247. This latter must I think be rejected, as there appears to be no sound basis whatever for his "probabilité propre." While there is a considerable range of values between ·002 and ·005, they may be described as being of the same order, and not incompatible with the differences in the data as to mortality, on which the methods applied are based, e.g. De Moivre's hypothesis, the actual reigns of European sovereigns, and the H^M mortality, which is a modern table.

If we take merely the relatively short series of German Emperors, who were *elected* sovereigns, the improbability is still more striking, i.e. ·00013. Not only is that material rather slender, but the process of election for the Emperors was very different from the process of election to the kingship of a small state; I should anticipate that the age of the Emperors at election was likely to be higher than that of the Kings of Rome. The average length of the pontificates is far less than that of the elected Emperors, and the few German Electors were likely to be influenced in the same manner, if not to the same extent, as the College of Cardinals.

Although there is, as Newton held, little doubt that the historians who put the total reigns of the first seven Kings of Rome at 257 years are untrustworthy, the methods by which a problem of this nature can be investigated seem to me far more important than the individual problem to which we have applied them. The actuarial method is the longer, but it will appeal to many as the better method, even when it has ultimately to become approximative. The method of small samples is shorter, but it needs in the first place the collection of data from reliable histories. It seems to me that much could be done to test the accuracy of twilight history by drawing statistics of events of a like class from records which are practically above criticism. For example the witnesses to charters, which grant extensive privileges to certain monastic houses, must quite frequently have been of fabulous ages at the dates when they are asserted to have been present at the granting of these charters. There are numerous documents in cartularies where the probability of forgery is a quantity definitely measurable; yet many such documents have been hitherto unhesitatingly used by historians.

ON THE USE AND INTERPRETATION OF CERTAIN TEST CRITERIA FOR PURPOSES OF STATISTICAL INFERENCE. PART II.

By J. NEYMAN, Ph.D. and E. S. PEARSON, D.Sc.

CONTENTS.

		PAGE
(1)	An Extension of the Definition of Likelihood	263
(2)	The Fundamental χ^2 Problem	265
(3)	The Test of Goodness of Fit	268
(4)	A Sampling Experiment	274
(5)	Summary of the Position	280
(6)	The Case of Two Samples	283
(7)	Application to Contingency Tables	284

(1) *An Extension of the Definition of Likelihood.*

In an earlier paper* we have endeavoured to emphasise the importance of placing in a logical sequence the stages of reasoning adopted in the solution of certain statistical problems, which may be termed problems of inference. In testing whether a given sample, Σ, is likely to have been drawn from a population Π, we have started from the simple principle that appears to be used in the judgments of ordinary life—that the degree of confidence placed in an hypothesis depends upon the relative probability or improbability of alternative hypotheses. From this point of view any criterion which is to assist in scaling the degree of confidence with which we accept or reject the hypothesis that Σ has been randomly drawn from Π should be one which decreases as the probability (defined in some definite manner) of alternative hypotheses becomes relatively greater. Now it is of course impossible in practice to scale the confidence with which we form a judgment with any single numerical criterion, partly because there will nearly always be present certain *à priori* conditions and limitations which cannot be expressed in exact terms. Yet though it may be impossible to bring the ideal situation into agreement with the real, some form of numerical measure is essential as a guide and control. In our previous paper we have made use of the criterion of likelihood. That there may be other forms of criteria or that this one can be interpreted in a different manner is very possible, but our object has been to find a single principle connecting the various sampling tests already in use, and one which could be extended to new problems.

* *Biometrika*, xx^A, pp. 175—240.

Hitherto no absolute value has been assigned to the measure of likelihood; if the chance of obtaining Σ from Π be C, and from Π' be C', then the *ratio of the likelihoods* of Π and Π' with regard to Σ has been defined as C/C'. This ratio of chances or frequencies is undoubtedly one which has a direct bearing on our judgments in ordinary life. In dealing, however, with continuous variation, for any unique sample Σ, both C and C' will be infinitesimally small. It is therefore necessary to take the chance of obtaining a sample whose variate values lie within certain fixed elements which are *independent* of Π and Π'. The value of the ratio C/C' will then depend upon the elements chosen. The ratio of the likelihoods has in this case been defined as the limiting value of C/C' as the elements tend to zero.

It is perhaps only easy to grasp the significance of the ratio in the case where C and C' are finite, as for example when dealing with grouped observations*. But if in giving the same interpretation to the limiting ratio we are making in a certain sense an assumption, so far we have been unable to discover any case where the use of this criterion leads to results which on intuitive grounds appear unreasonable.

It was in this form of the ratio of likelihoods that the idea was introduced by R. A. Fisher†, and it was so used in our previous paper. It is, however, possible and perhaps desirable to give a more precise definition of the term. The populations Π, Π', etc. which are considered as possible sources of origin for Σ must belong to some set or universe of populations, Ω, which can generally be specified; further, Ω will contain, in all cases likely to be met in practice, a certain population Π ($\Omega_{\max.}$) for which the chance of drawing Σ is a maximum, say $C = C$ ($\Omega_{\max.}$). We may now define the likelihood of the hypothesis that Σ has been drawn from Π as

$$\lambda = C/C\,(\Omega_{\max.}) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1).$$

Under these conditions the maximum likelihood becomes unity. The likelihood, λ, is of course defined with regard to a definite set of populations, and will change if Ω be changed.

The hypothesis that Σ has been drawn from an exactly defined population Π may be termed a simple hypothesis, while we may term the hypothesis that Σ has been drawn from some unspecified population belonging to a clearly defined sub-set ω of the set Ω, a composite hypothesis. For example, if Ω represents all possible normal frequency distributions

$$y = \frac{1}{\sqrt{2\pi}\sigma}\, e^{-\frac{1}{2}\frac{(x-a)^2}{\sigma^2}},$$

a simple hypothesis is that Σ has been sampled from a population for which $a = a_1$, $\sigma = \sigma_1$, while a composite hypothesis is that in the population $a = a_1$ but σ may have any value.

We can now define the likelihood of a composite hypothesis. Let the population of the sub-set ω for which the chance of occurrence of Σ is a maximum be

* The ratio is not here final in the sense that the limiting ratio is, since C and C' will change with the arrangement of grouping.

† *Phil. Trans.* Vol. 222, A, p. 326.

$\Pi(\omega_{max.})$, and let this chance be $C(\omega_{max.})$. Then we define the likelihood of the composite hypothesis that Σ has been sampled from some member of the sub-set ω of the set Ω to be the ratio*,

$$\lambda_1 = C(\omega_{max.})/C(\Omega_{max.}) \quad \ldots\ldots(2)$$

"Student's" Problem consists in testing such a composite hypothesis, and this definition leads at once to the expression (xxv) of our first paper, namely,

$$\lambda_1 = (1+z^2)^{-\frac{n}{2}},$$

so that the contours of constant likelihood are the z-contours.

The value of this criterion in the case of testing composite hypotheses will perhaps be questioned. It may be argued that it is impossible to estimate the probability of such a hypothesis without a knowledge of the relative *à priori* probabilities of the constituent simple hypotheses. But in general it is quite impossible even to attempt to express our *à priori* knowledge in exact terms. Can we then get no further? Certainly we are faced with a problem whose solution cannot take the same form as that which is possible when a simple hypothesis is tested, yet we are inclined to think that this ratio of maximum chances or frequencies of occurrence provides perhaps the best numerical measure that can be found under the circumstances to guide our judgment.

As for λ, so in the case of λ_1 a knowledge of the ratio alone is not however sufficient. It has been pointed out that the errors of judgment which must inevitably occur will be of two kinds, (1) we sometimes reject the hypothesis when it is in fact true, and (2) we sometimes accept it when Σ has been drawn from some population not belonging to the sub-set ω. The form of the criterion λ_1 has been chosen to minimise the effect of (2), but it is necessary to know also the probability distribution of λ_1 in sampling from the members of ω in order to control the source of error (1). In the cases which we have so far come across this distribution has possessed the essential property of being the same or approximately the same whatever member of ω may have been sampled.

(2) *The Fundamental χ^2 Problem.*

We have indicated† in the case of a single continuous variable the geometrical relation in multiple space between the point density

$$D = f(x_1)f(x_2)\ldots f(x_N) \quad \ldots\ldots(3)$$

associated with a sample $x_1, x_2 \ldots x_N$ from a population following the distribution law $y = f(x)$, and the term of the multinomial

$$C = \frac{N!}{n_1! \ldots n_k!}(p_1)^{n_1} \ldots (p_k)^{n_k} \quad \ldots\ldots(4)$$

* This will be as before a limiting ratio if we are dealing with continuous variation and observations which are not grouped.

† *Biometrika*, xx^A, p. 185.

representing the chance of drawing a sample with group frequencies $n_1, n_2 \ldots n_k$ from a population in which the corresponding group proportions are $p_1, p_2 \ldots p_k$. The expression (4) is, however, of very general application, for in using it we are not limited to problems of a single variable, to equal grouping ranges or even to quantitative variables at all. Further, instead of using the N-dimensioned space previously considered, it is now possible to represent the problem in a space of k dimensions, in which the variables are the sample group frequencies $n_1, n_2 \ldots n_k$, while the relation

$$n_1 + n_2 + \ldots + n_k = N \quad \ldots\ldots(5)$$

representing a prime in this space, shows that the sample point P is constrained to move in a space of $k-1$ dimensions.

As the quantities n_s must increase by integral values, we are in fact only concerned with a finite number of discrete points in the space, with each of which is associated a value of C. But if the size of the sample be large so that there exists a very great variety of possible sample points, and if none of the quantities $p_s N$ are too small, it becomes possible to represent with good approximation the sum of discrete values of C throughout a given region of the field by the integral of a continuous density function throughout that region. This function we shall now consider. Write

$$\tilde{m}_s = p_s N, \text{ and } \delta n_s = n_s - \tilde{m}_s \qquad (s = 1, 2 \ldots k) \ldots(6),$$

where of course

$$\tilde{m}_1 + \tilde{m}_2 + \ldots + \tilde{m}_k = N \quad \ldots\ldots(7).$$

Then (4) may be written

$$C = \frac{N!}{N^N} \frac{(\tilde{m}_1)^{n_1}}{n_1!} \ldots \frac{(\tilde{m}_k)^{n_k}}{n_k!} \quad \ldots\ldots(8).$$

If the values of all the group frequencies n_s are large enough to justify, as an approximation, the neglect of the terms $1/12n_s$ and those which follow in Stirling's Expansion for $(n_s)!$, it is known that (8) may be written*

$$C = \frac{1}{(\sqrt{2\pi})^{k-1}} \sqrt{\frac{N}{\tilde{m}_1 \ldots \tilde{m}_k}}\, e^{-\frac{1}{2}\sum_{s=1}^{k} \left\{ \frac{\delta n_s^2}{\tilde{m}_s} - \frac{1}{3}\frac{\delta n_s^3}{\tilde{m}_s^2} + \frac{\delta n_s}{\tilde{m}_s} - \frac{1}{2}\frac{\delta n_s^2}{\tilde{m}_s^2} + \ldots \right\}} \quad \ldots\ldots(9).$$

All the terms except the first in the power of the exponential tend to zero as the $\tilde{m}_s$ increase. We take, therefore, as an approximation to C, the continuous density field represented by

$$D = D_0 e^{-\frac{1}{2}\chi^2} \quad \ldots\ldots(10),$$

where

$$\chi^2 = \sum_{s=1}^{k} \left(\frac{\delta n_s^2}{\tilde{m}_s} \right) = \sum_{s=1}^{k} \frac{(n_s - \tilde{m}_s)^2}{\tilde{m}_s} \quad \ldots\ldots(11).$$

The assumptions involved in obtaining (10) from the multinomial are the same as those of the original method†, which reached the same result starting from the multivariate normal correlation surface of the sample group frequencies.

* See for example E. Borel, *Traité du Calcul des Probabilités*, Tome 1, Fasc. 1, pp. 35—36.

† K. Pearson, *Phil. Mag.* July 1900, pp. 157—175.

The next problem is to consider the contours which should be used in testing the hypothesis that Σ has been drawn from Π. We shall use the contours of likelihood. In the earlier paper, when dealing with continuous variation in small samples, it was necessary to limit the set, Ω, of possible populations (as defined on p. 264 above) to normal distributions, to rectangular distributions, etc. This is no longer necessary, for Ω may now be taken as the set of all possible populations classed into k frequency groups. For such populations $C\,(\Omega_{\max.})$ is obtained by making

$$C' = N!\,(p_1')^{n_1} \dots (p_k')^{n_k}/(n_1! \dots n_k!)$$

a maximum subject to $p_1' + p_2' + \dots + p_k' = 1$. The solution is

$$p_s' = n_s/N, \qquad (s = 1, 2 \dots k).$$

Consequently the likelihood of this simple hypothesis becomes

$$\lambda = \frac{C}{C\,(\Omega_{\max.})} = \left(\frac{\widetilde{m}_1}{n_1}\right)^{n_1} \dots \left(\frac{\widetilde{m}_k}{n_k}\right)^{n_k} \qquad \dots\dots\dots\dots(12).$$

But
$$\log \lambda = -\,n_1 \log\left(1 + \frac{\delta n_1}{\widetilde{m}_1}\right) - \dots - n_k \log\left(1 + \frac{\delta n_k}{\widetilde{m}_k}\right)$$
$$= -\sum_{s=1}^{k} \left\{(\widetilde{m}_s + \delta n_s)\left(\frac{\delta n_s}{\widetilde{m}_s} - \tfrac{1}{2}\left(\frac{\delta n_s}{\widetilde{m}_s}\right)^2 + \tfrac{1}{3}\left(\frac{\delta n_s}{\widetilde{m}_s}\right)^3 - \dots\right)\right\}$$
$$= -\sum_{s=1}^{k} \left\{\delta n_s + \tfrac{1}{2}\frac{\delta n_s^{\,2}}{\widetilde{m}_s} - \tfrac{1}{6}\frac{\delta n_s^{\,3}}{\widetilde{m}_s^{\,2}} + \dots\right\},$$

and since $S(\delta n_s) = 0$, we find that

$$\lambda = e^{-\frac{1}{2}\sum_{s=1}^{k}\left\{\frac{\delta n_s^{\,2}}{\widetilde{m}_s} - \frac{1}{3}\frac{\delta n_s^{\,3}}{\widetilde{m}_s^{\,2}} + \dots\right\}} \qquad \dots\dots\dots\dots(13),$$

or approximately
$$\lambda = e^{-\frac{1}{2}\chi^2} \qquad \dots\dots\dots\dots(14).$$

It follows that the levels of constant likelihood, λ, among the discrete sample points will correspond approximately to the contours of χ in the density field represented by (10)*, and we may take as the chance of obtaining a sample from Π with $\lambda \leqslant \lambda_0$ that of obtaining $\chi \geqslant \chi_0$, or

$$P = \frac{\int_{\chi_0}^{\infty} \chi^{k-2} e^{-\frac{1}{2}\chi^2} d\chi}{\int_{0}^{\infty} \chi^{k-2} e^{-\frac{1}{2}\chi^2} d\chi} \qquad \dots\dots\dots\dots(15).$$

From this point of view we give another interpretation to the ordinary (P, χ^2) Test; we accept λ rather than χ as the fundamental criterion by which to measure our confidence in the hypothesis that Σ has been sampled from Π. It would of course be unreasonable to accept as the most probable alternative the population of maximum likelihood, namely, that in which $p_s = n_s/N$, but these proportions point to a *direction* in which there will lie a widening field of possible alternatives as the likelihood diminishes. And this is really the idea which underlies the use of the λ contours.

* The relation between χ and the method of likelihood has been discussed by R. A. Fisher, *Phil. Trans.* Vol. 222, A, pp. 356—358; *Jour. Roy. Stat. Soc.* Vol. LXXXVII. pp. 442—450.

(3) *The Test of Goodness of Fit.*

In the preceding section the standard χ^2 result has been shown to follow from the use of the criterion of likelihood; it was there supposed that the population Π could be exactly specified, that is to say, we were dealing with a simple hypothesis. We have now another kind of problem in which a theoretical law of frequency of given type, but whose exact form depends upon certain undetermined parameters, is fitted to a series of observations, the value obtained for the parameters depending on the method of fitting. Some measure is then required of the adequacy of this fit.

If there be c of these parameters or constants to be determined*, then the proportional frequency in the sth group into which the population is classed may be expressed by a relation

$$p_s = f(s;\ \alpha_1, \alpha_2 \dots \alpha_c) \quad \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots(16),$$

where the form of f is known, but $\alpha_1, \alpha_2 \dots \alpha_c$ are to be determined from the sample. Using some appropriate method of fitting, (16) will give a series of frequencies

$$m_s = Np_s = Nf(s;\ \alpha_1, \alpha_2 \dots \alpha_c) \qquad (s = 1, 2 \dots k) \dots(17)$$

with which to compare the observed frequencies $n_1, n_2 \dots n_k$. It is quite possible that all we may then require is an answer to the question: *Is it probable that the observed sample would have arisen in random sampling from a population whose group proportions are actually the values* $p_s = m_s/N$ *of* (17) *determined by fitting* (16) *to the observations?* This question is answered with the help of the ordinary (P, χ^2) Test, putting $\chi = \chi_1$, where

$$\chi_1^2 = \sum_{s=1}^{k} \frac{(n_s - m_s)^2}{m_s} \quad \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots(18),$$

and the Tables are entered with $n' = k$. We are testing a simple hypothesis.

On the other hand, we may attempt to answer a rather different question: *Is it probable that the observed sample would have occurred in random sampling from a population whose law of frequency is defined by* (16)*?* These two questions are not the same; how far an answer to the first can be considered as satisfying the second will be discussed later. In recent papers† R. A. Fisher has given results showing that under certain conditions the second question can be answered by using the (P, χ^2) Test with $\chi^2 = \chi_1^2$ as in (18) and $n' = k - c$. Partly because we shall approach the problem from a somewhat different point of view and partly because we believe that some may have difficulty in following Dr Fisher's very condensed reasoning we propose to go into this question at some length. We shall also deduce the correlation surface of the χ of (11) and the χ_1 of (18), and test the theoretical formulae by applying them to the results of experimental sampling.

If the differences between the sample group frequencies, n_s, and the corresponding reduced population frequencies, $\tilde{m}_s$, or $x_s = n_s - \tilde{m}_s$, $(s = 1, \dots k)$, are taken

* c is the minimum number of parameters required to express the law.

† *Jour. Roy. Stat. Soc.* LXXXV. pp. 87—94, and LXXXVII. pp. 442—450.

as variables, the problem may be represented in a k-dimensioned space with axes Ox_1, $Ox_2 \ldots Ox_k$, in which the sample points are constrained to lie on the prime

$$x_1 + x_2 + \ldots + x_k = 0 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (19),$$

or in $k-1$ dimensions. The chance of drawing from the population a sample lying in a given region of this space may then be represented approximately, as has been seen, by the integral throughout this region of the continuous density function

$$D = D_0 e^{-\frac{1}{2}\chi^2} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (10\ bis),$$

where we may now write

$$\frac{x_1^2}{\tilde{m}_1} + \frac{x_2^2}{\tilde{m}_2} + \ldots + \frac{x_k^2}{\tilde{m}_k} = \chi^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (20).$$

If now we fit to the sample a frequency law of given form depending on c parameters, the resulting frequencies m_1, $m_2 \ldots m_k$ will be represented by a point

$$X_1 = m_1 - \tilde{m}_1, \ldots X_k = m_k - \tilde{m}_k,$$

which must lie on what may be termed the population locus, or

$$\begin{cases} X_s = f_s(\alpha_1, \alpha_2 \ldots \alpha_c) \qquad (s = 1, 2 \ldots k-1) \quad \ldots (21\ a), \\ X_1 + X_2 + \ldots + X_k = 0 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (21\ b). \end{cases}$$

For a given sample $(x_1, \ldots x_k)$, the point $(X_1, \ldots X_k)$ chosen will depend upon the method of fitting employed. The value of χ_1^2 calculated from the fitted population frequencies is

$$\frac{(x_1 - X_1)^2}{\tilde{m}_1 + X_1} + \ldots + \frac{(x_k - X_k)^2}{\tilde{m}_k + X_k} = \chi_1^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (22).$$

This expression represents a hyperellipsoid centred at $(X_1, \ldots X_k)$. Suppose now that the method of fitting is one which determines $(X_1, \ldots X_k)$, subject to equations (21), so that χ_1^2 is a minimum. Then it follows that the locus of the sample points for which χ_1^2 is constant is the envelope of the hyperellipsoids (22) whose centres are constrained to move on the population locus (21). Under these conditions, therefore, the chance of obtaining a value of χ_1 greater let us say than γ will be equal to the integral of D taken outside the envelope of the hyperellipsoids for which $\chi_1 = \gamma$. A general solution for this integration would be extremely difficult to obtain; in the present paper we shall, however,

(a) Show how a solution may be obtained under certain simplifying assumptions.

(b) Give in an Appendix what we believe to be a rigorous proof that the approximation (a) is the limiting form for large samples under very general conditions.

(c) Describe certain experimental results which suggest that this limiting solution will in fact represent adequately the situation met with in common practice.

The assumptions required for the solution (a) are as follows:

(1) That, as before, the distribution of samples from the population Π can be adequately represented by the continuous density function D of (10).

(2) That the point $(X_1, \ldots X_k)$ is sufficiently close to the origin—or the fitted m_s near enough to the true $\tilde{m}_s$—for the linear transformation by which the hyperellipsoids (20) are turned into concentric hyperspheres of radius χ*, or

$$x_1'^2 + x_2'^2 + \ldots + x_k'^2 = \chi^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(23)$$

to change also the hyperellipsoids (22) approximately into hyperspheres of radius χ_1, or

$$(x_1' - X_1')^2 + \ldots + (x_k' - X_k')^2 = \chi_1^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots(24).$$

In other words, we assume that at any rate within the region of significant density, the terms $X_1, \ldots X_k$ in the denominators of (22) are negligible compared with $\tilde{m}_1, \ldots \tilde{m}_k$.

(3) If the c constants α be eliminated from (21 a), we shall be left with $k-c-1$ equations of form

$$F_t(X_1, \ldots X_k) = 0 \qquad (t = 1, 2 \ldots k-c-1) \quad \ldots(25).$$

Since $X_s = m_s - \tilde{m}_s$ is small these equations may generally be expanded by Taylor's Theorem in powers of the X's. It will be assumed that in the neighbourhood of the origin it is adequate to take the first order terms only in these expansions, and to represent the population locus, instead of by (21), by the $k-c$ primes

$$\begin{cases} a_{1t}X_1 + \ldots + a_{kt}X_k = 0 \qquad (t = 1, 2 \ldots k-c-1) \quad \ldots(26\,a), \\ \quad X_1 + \ldots + \quad X_k = 0 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(26\,b). \end{cases}$$

If these three assumptions are justified the same linear transformation will change the hyperellipsoids (22) as well as the hyperellipsoids (20) into hyperspheres, and will leave the equations (26) as a series of primes. The effect of the conditions (19) and (26 b) is to limit the field to a space of $k-1$ dimensions. It follows that the problem reduces to that of integrating D outside the envelope traced out in the $(k-1)$-dimensioned space defined by (19), by a hypersphere of radius χ_1 whose centre is constrained to move in this space on the $k-c-1$ primes (26 a)†.

We cannot visualise the relation of the hypersurfaces and envelopes beyond the case where $k=4$, but in this case the relationship can be represented in three dimensions; the contours of χ are ordinary ellipsoidal surfaces in three dimensions surrounding the population point O. If $c=1$, the point Q representing the population obtained by fitting to the sample is constrained to lie on a locus of one dimension, i.e. a curve, which by the assumption (3) above we suppose to lie in the neighbourhood of O on the intersection of two planes, or to be represented by a straight line. After the transformation, the locus of constant χ_1 is approximately the envelope of spheres of radius χ_1 whose centres move on a straight line, that is to say, it is a cylinder. If $c=2$, then the locus of Q is two-dimensioned, and we assume that near O it can be represented by a plane. The locus of constant χ_1 is now approximately the envelope of spheres whose centres move on a plane, that is to say, consists of two parallel planes at a distance $2\chi_1$ apart.

* x' and X' represent the coordinates after transformation.

† More correctly, we shall be dealing with (19) and (26 a) when transformed into terms of x' and X'.

To obtain the distribution of χ_1 we shall consider the solution of the following general lemma.

Lemma. A point O in a space of n-dimensions is surrounded by a density field

$$D = D_0 e^{-\frac{1}{2}z^2} \quad \text{.....(27),}$$

where

$$x_1^2 + x_2^2 + \ldots + x_n^2 = z^2 \quad \text{.....(28).}$$

The chance of a point P falling into any region of this space is measured by the integral of D taken throughout that region. A point Q is constrained to move upon a locus defined by the intersection of s primes,

$$a_{1t}x_1 + a_{2t}x_2 + \ldots + a_{nt}x_n = 0 \qquad (t = 1, 2 \ldots s) \ldots (29).$$

Then the problem is to find the chance that the distance, u, between P and the nearest possible position of Q lies within the limits $u \pm \frac{1}{2}du$, du being an infinitesimal element.

Let P be the point $(x_1, x_2 \ldots x_n)$ and Q the point $(X_1, X_2 \ldots X_n)$. Then

$$(x_1 - X_1)^2 + \ldots + (x_n - X_n)^2 = u^2 \quad \text{.....(30),}$$

and

$$a_{1t}X_1 + a_{2t}X_2 + \ldots + a_{nt}X_n = 0 \qquad (t = 1, 2 \ldots s) \ldots (31).$$

Let Figure 1 represent the section of the n-dimensioned space by the plane POQ; call $OQ = w$.

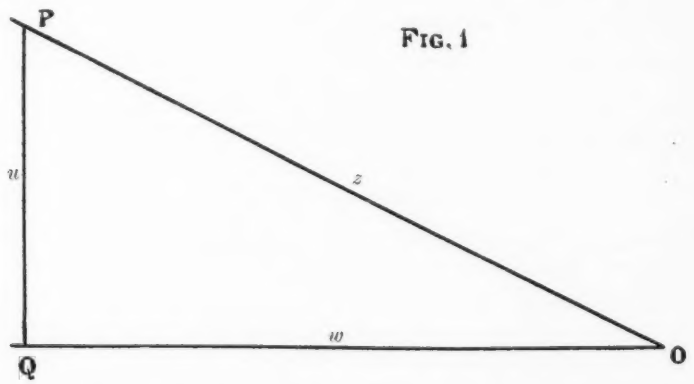


FIG. 1

For a given P, Q is determined in the following manner.

By solving the s equations (31) it is possible to express s of the quantities $X_1, \ldots X_n$ as linear functions of the remaining $n - s$, say $X_1, X_2 \ldots X_{n-s}$; these values may be substituted into (30) and will give on minimising u^2, $n - s$ linear relations of form

$$\frac{\partial(u^2)}{\partial X_t} = 0 \qquad (t = 1, 2 \ldots n - s) \quad \text{.....(32).}$$

These relations (32) will determine the unique point Q on (29) for which $PQ = u$ is a minimum. It follows conversely that for a given point Q the locus of points P for which

(1) $PQ = u =$ constant,

(2) Q is the nearest point to P lying on (29)

is the intersection of the $(n-1)$-dimensioned hypersphere* centred at Q, radius u with the $(n-s)$ primes (32), which may now be regarded as linear equations in the x's. That is to say, the locus is an $(s-1)$-dimensioned hypersphere lying at right angles to (29). It is also clear that PQ must be at right angles to OQ, seeing that OQ must lie in the locus (29) and that PQ is the minimum distance from P to this locus. Hence for a fixed Q and constant u, $z^2 = u^2 + w^2$ and is constant, so that this locus of P for fixed Q lies on a single one of the hyperspherical density contours defined by (27) and (28).

Consequently the chance of drawing a sample for which

(1) $u + \frac{1}{2}du \geqslant PQ \geqslant u - \frac{1}{2}du$,

(2) Q is at a fixed position on (29),

(3) $u = PQ$ is a minimum

$$\propto u^{s-1}du\, D_0 e^{-\frac{1}{2}z^2},$$

$$\propto u^{s-1}e^{-\frac{1}{2}u^2}du \times e^{-\frac{1}{2}w^2}.$$

To complete the solution it is necessary to integrate this expression for all possible positions of Q on (29), remembering that for a given P, Q is uniquely determined; that is to say, we must take $\int e^{-\frac{1}{2}w^2}dv$ over the hypervolume defined by (29). The result will be clearly independent of u, and it follows that the chance of obtaining a point P for which PQ lies in the limits $u \pm \frac{1}{2}du$

$$\propto u^{s-1}e^{-\frac{1}{2}u^2}du,$$

or
$$f(u)\,du = c_0 u^{s-1}e^{-\frac{1}{2}u^2}du \quad \text{.....(33).}$$

A further problem is to find the chance of obtaining a point P for which u and z lie within the limits $u \pm \frac{1}{2}du$, and $z \pm \frac{1}{2}dz$.

For this we take the expression

$$u^{s-1}e^{-\frac{1}{2}u^2}du \int e^{-\frac{1}{2}w^2}dv \text{.....(34),}$$

where the integral is now to be taken over that portion only of the hypervolume (29) for which $w^2 = z^2 - u^2$. Now (29) cuts the hypersphere (28) in an $(n-s-1)$-dimensioned hypersphere, and consequently (34) is proportional to

$$u^{s-1}e^{-\frac{1}{2}u^2}w^{n-s-1}e^{-\frac{1}{2}w^2}du\,dw.$$

* We shall speak of the locus represented by an equation such as (28) as an $(n-1)$-dimensioned hypersphere; for if $n=3$ we have the surface of an ordinary sphere in 3-dimensioned space, which is a 2-dimensioned locus. Similarly if $n=2$ the locus is a circle, or a figure of 1-dimension lying in a plane.

Changing the variables from u and w to u and z, since $z=\sqrt{u^2+w^2}$, we have $du\,dw = du\,dz\,.\,z/w$, and consequently

$$\phi(u,z)\,du\,dz \propto u^{s-1}e^{-\frac{1}{2}u^2}w^{n-s-2}e^{-\frac{1}{2}w^2}z\,du\,dz$$

$$= c_0' u^{s-1} z\,(z^2-u^2)^{\frac{n-s-2}{2}} e^{-\frac{1}{2}z^2}du\,dz \quad \ldots\ldots\ldots\ldots\ldots(35).$$

These results may be applied at once to the (χ^2, χ_1^2) problem as it has been simplified by the assumptions of p. 270. We must write $z=\chi$, $u=\chi_1$ and as the problem is reduced to $k-1$ dimensions by condition (19), $n=k-1$. Also the s primes (29) correspond to the $k-c-1$ equations (26 a), so that $s=k-c-1$.

It follows from (33) that

$$f(\chi_1)\,d\chi_1 = c_0 \chi_1^{k-c-2} e^{-\frac{1}{2}\chi_1^2} d\chi_1 \quad \ldots\ldots\ldots\ldots\ldots(36),$$

and from (35) that

$$\phi(\chi, \chi_1)\,d\chi\,d\chi_1 = c_0' \chi_1^{k-c-2} \chi\,(\chi^2-\chi_1^2)^{\frac{c-2}{2}} e^{-\frac{1}{2}\chi^2} d\chi\,d\chi_1 \quad \ldots\ldots\ldots(37).$$

Equation (36) represents the distribution referred to above which has been given by Fisher; it shows that if the assumptions involved are justified the distribution of minimum χ_1 is that of the ordinary χ obtained by using the true population frequencies, if the group number k be reduced by the number of constants in the population law which are determined by fitting to the sample. The relation (37) is we believe novel. Both these results assume a more convenient form if we take $\psi=\chi^2$ and $\psi_1=\chi_1^2$ as variables. We have, then, for the distributions of ψ and ψ_1 the Type III curves

$$y = y_0 \psi^{\frac{k-3}{2}} e^{-\frac{1}{2}\psi} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(38),$$

$$y = y_0' \psi_1^{\frac{k-c-3}{2}} e^{-\frac{1}{2}\psi_1} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(39),$$

and these are the margins of the correlation surface

$$z = z_0 \psi_1^{\frac{k-c-3}{2}} (\psi-\psi_1)^{\frac{c-2}{2}} e^{-\frac{1}{2}\psi} \quad \ldots\ldots\ldots\ldots\ldots(40).$$

This surface (31) has the following properties:

(a) It is limited by the line $\psi=\psi_1$, i.e. $\psi \geqslant \psi_1$.

(b) For a given ψ the array of ψ_1 is a Type I curve for which it is easy to show that

$$\text{Mean } \psi_1 = \psi\,(k-c-1)/(k-1),$$

or the regression of ψ_1 on ψ is linear.

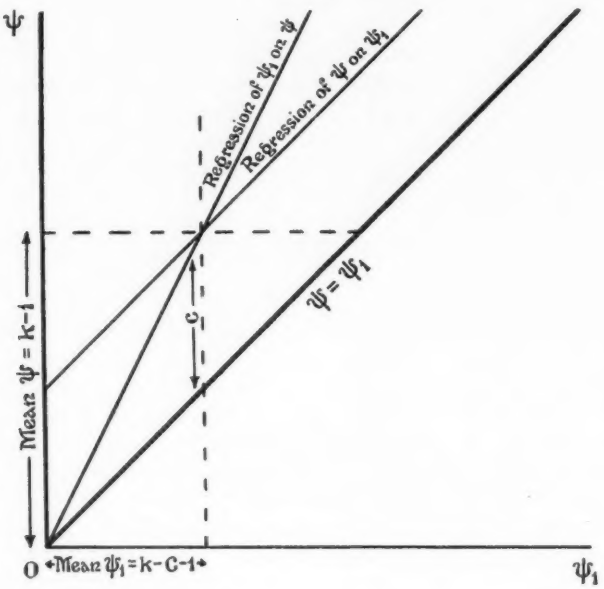
(c) For a given ψ_1 the array of ψ is a Type III curve with start at the point $\psi=\psi_1$, and with mean at a distance c and mode at a distance $c-2$ from this point. The regression of ψ on ψ_1 is therefore linear.

(d) Since $\sigma_\psi = \sqrt{2\,(k-1)}$ and $\sigma_{\psi_1} = \sqrt{2\,(k-c-1)}$, we obtain from either regression line the value for the correlation between ψ and ψ to be

$$r_{\psi\psi_1} = \sqrt{\frac{k-c-1}{k-1}}.$$

The correlation plane is represented diagrammatically in Figure 2.

FIG. 2



(4) *A Sampling Experiment.*

In carrying out this piece of work the authors are very much indebted to Miss M. Page, without whose help both in sampling and in the lengthy computation the experiment could never have been undertaken.

The population law was chosen to follow a cubic ($c=3$), and the area under the curve was broken up into 8 groups ($k=8$). When reduced to a total of $N=200$, the population law was

$$y = 25 + \tfrac{45}{8}x - \tfrac{15}{128}x^3 \quad \text{.....(41)},$$

the range being from $x=-4$ to $x=+4$. The curve which has zero differential coefficients at the terminals, and a point of inflexion at the centre, is drawn approximately in Figure 3.

The reduced group frequencies are (to one decimal place)

$\tilde{m}_1$	$\tilde{m}_2$	$\tilde{m}_3$	$\tilde{m}_4$	$\tilde{m}_5$	$\tilde{m}_6$	$\tilde{m}_7$	$\tilde{m}_8$	N
10·4	12·8	17·0	22·2	27·8	33·0	37·2	39·6	200

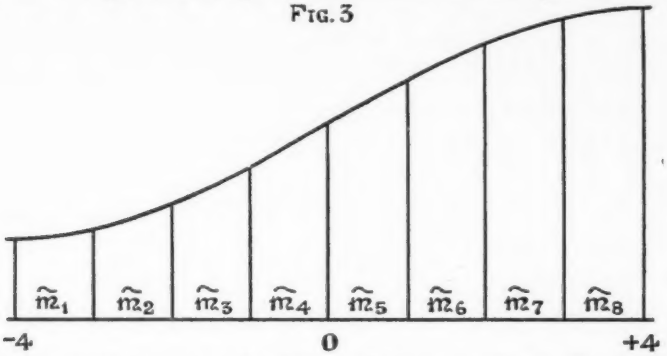
208 random samples of 200 were then taken, using Tippett's Random Numbers*. To avoid any possible unevenness in these numbers, the scale was changed four times during the sampling; for example, with "Scale A" the distribution of the 10,000 numbers into 8 groups was as follows:

* *Tracts for Computers*, No. xv.

Group	1	2	3	4
No.'s in book	0000—0521	0522—1163	1164—2013	2014—3124
Group	5	6	7	8
No.'s in book	3125—4513	4514—6163	6164—8021	8022—9999

As it would only seem possible to fit a cubic by the methods of Minimum χ_1^2 or Maximum Likelihood by resorting to successive approximations, fitting by moments was adopted. We shall not, therefore, obtain the distribution of minimum χ_1^2, but the result should be of the more value because it will show how far the distribution of a χ_1^2 obtained by a common method of fitting used in ordinary

FIG. 3



practice follows the theoretical distribution of minimum χ_1^2—or rather the limiting distribution of this quantity expected in very large samples.

It was necessary to find the constants a_0, a_1, a_2, a_3 by fitting a cubic $y = a_0 + a_1x + a_2x^2 + a_3x^3$ to each sample. The process employed consisted in equating the zero, first, second and third order moments of the theoretical group frequencies $m_1, m_2, \ldots m_8$* to the corresponding moments of the observed frequencies $n_1, n_2, \ldots n_8$. (No corrections are in this case necessary as the moments on both sides of the equations are grouped.) The four resulting equations when solved give a_0, a_1, a_2 and a_3 as linear functions of $n_1, \ldots n_8$, and these values when substituted into the expressions for $m_1, \ldots m_8$ (of which that for m_1 is given in the footnote) result in the following eight relations expressing $m_1, \ldots m_8$ as linear functions of the observed sample group frequencies:

$$\left\{\begin{aligned}
m_1 = {} & \cdot 8939n_1 + \cdot 2424n_2 - \cdot 0606n_3 - \cdot 1212n_4 - \cdot 0454n_5 + \cdot 0606n_6 \\
& + \cdot 0909n_7 - \cdot 0606n_8, \\
m_2 = {} & \cdot 2424n_1 + \cdot 3745n_2 + \cdot 3290n_3 + \cdot 1818n_4 + \cdot 0087n_5 - \cdot 1147n_6 \\
& - \cdot 1126n_7 + \cdot 0909n_8, \\
m_3 = {} & - \cdot 0606n_1 + \cdot 3290n_2 + \cdot 4177n_3 + \cdot 3117n_4 + \cdot 1168n_5 - \cdot 0606n_6 \\
& - \cdot 1147n_7 + \cdot 0606n_8, \\
m_4 = {} & - \cdot 1212n_1 + \cdot 1818n_2 + \cdot 3117n_3 + \cdot 3139n_4 + \cdot 2338n_5 + \cdot 1169n_6 \\
& + \cdot 0087n_7 - \cdot 0455n_8
\end{aligned}\right. \qquad \ldots\ldots(42).$$

* For example $m_1 = \int_{-4}^{-3} (a_0 + a_1x + a_2x^2 + a_3x^3)\,dx = a_0 - \cdot\frac{1}{2}a_1 + \frac{3}{2}a_2 - 1\frac{1}{4}a_3.$

The expressions for m_5, m_6, m_7 and m_8 are obtained by reversing the order of the multipliers in the expressions for m_4, m_3, m_2, m_1 respectively. For example,

$$m_7 = \cdot0909n_1 - \cdot1126n_2 - \cdot1147n_3 + \cdot0087n_4 + \cdot1818n_5 + \cdot3290n_6 + \cdot3745n_7 + \cdot2424n_8.$$

Using these expressions the fitted population frequencies, $m_1, \ldots m_8$, could each be computed by a continuous operation on the calculating machine, for each of the 208 samples. And from these $\chi^2 = \sum_1^8 (n_s - \tilde{m}_s)^2/\tilde{m}_s$ and $\chi_1^2 = \sum_1^8 (n_s - m_s)^2/m_s$ were calculated. The resulting (χ^2, χ_1^2) distribution is shown in Table I. Consider now, (*a*) the χ^2 margin, (*b*) the χ_1^2 margin, (*c*) the correlation and regression, and (*d*) the distribution in the arrays of constant χ_1^2.

(*a*) *The χ^2 Margin.*

Since $k = 8$, the theoretical distribution of $\psi = \chi^2$ is $y = y_0 \psi^{\frac{5}{2}} e^{-\frac{1}{2}\psi}$, a Type III curve with

$$\begin{cases} \text{Mean } \psi = k - 1 = \ 7{\cdot}0000 \text{ and standard error for 208 samples of } 0{\cdot}2594, \\ \sigma_\psi^2 = 2(k-1) = 14{\cdot}0000 \qquad \text{,,} \qquad \text{,,} \qquad \text{,,} \qquad 1{\cdot}8708^*. \end{cases}$$

The observations give Mean $\psi = 6{\cdot}6767$ and $\sigma_\psi^2 = 12{\cdot}8806$, the first differing from theory by a little more and the latter by a little less than the standard error.

Applying the Goodness of Fit Test to the observations and the corresponding Type III frequencies obtained from the (P, χ^2) Tables, it is found that with 12 groups ($n' = 12$), $\chi^2 = 8{\cdot}85$ and $P = {\cdot}64$.

(*b*) *The χ_1^2 Margin.*

Since $k = 8$ and $c = 3$, the fourth constant of the cubic being of course simply that required to make the totals of observation and theory agree, the distribution of $\psi_1 = \textit{minimum } \chi_1^2$ should be $y = y_0 \psi_1 e^{-\frac{1}{2}\psi_1}$, a Type III curve with

$$\begin{cases} \text{Mean } \psi_1 = k - c - 1 = 4{\cdot}0000 \text{ and standard error for 208 samples of } 0{\cdot}1961, \\ \sigma_{\psi_1}^2 = 2(k-c-1) = 8{\cdot}0000 \qquad \text{,,} \qquad \text{,,} \qquad \text{,,} \qquad 1{\cdot}2403^*. \end{cases}$$

The observations give Mean $\psi_1 = 4{\cdot}0421$ and $\sigma_{\psi_1}^2 = 7{\cdot}6012$, both differing from the theoretical values for *minimum* ψ_1 by considerably less than their standard errors. Applying the Goodness of Fit Test, taking the theoretical frequencies from the (P, χ^2) Tables, it is found that with 9 groups $\chi^2 = 8{\cdot}18$ and $P = {\cdot}42$.

(*c*) *The Correlation and Regression.*

From (40) it follows that the theoretical correlation surface for $\chi^2 = \psi$ and minimum $\chi_1^2 = \psi_1$ is

$$z = z_0 \psi_1 (\psi - \psi_1)^{\frac{1}{2}} e^{-\frac{1}{2}\psi}.$$

* The standard errors used for the variance are $\sigma_\psi^2 \sqrt{(\beta_2 - 1)/N}$, where $N = 208$, the number of samples, and β_2 for a theoretical χ^2 distribution is $3 + 12/(k-1)$, k in the three sections (1), (2) and (4) taking the values of 8, 5 and 4 respectively.

TABLE I.

TABLE I.

Experimental Distribution of χ^2 and χ_1^2 in 208 Samples of 200.

Central Values of χ_1^2.

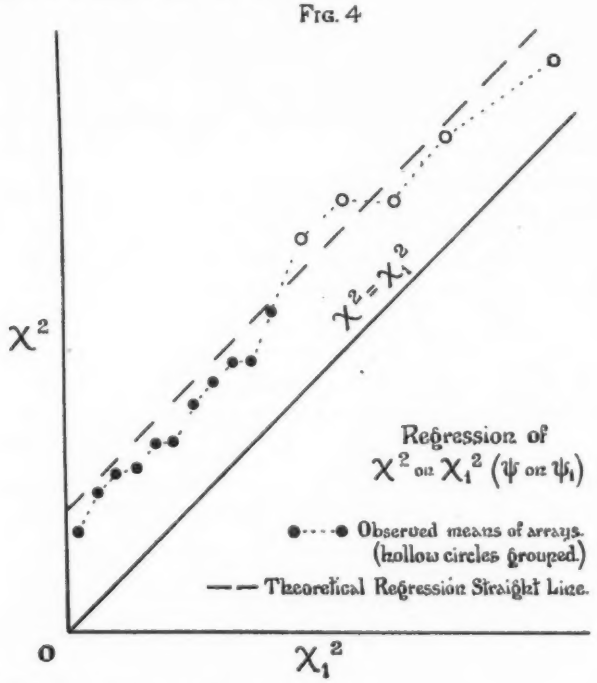
Central Values of χ^2.	0·25	0·75	1·25	1·75	2·25	2·75	3·25	3·75	4·25	4·75	5·25	5·75	6·25	6·75	7·25	7·75	8·25	8·75	9·25	9·75	10·25	10·75	11·25	11·75	12·25	12·75	—	—	16·25	Totals
19·75	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	—	—	—	—	1
19·25	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
18·75	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
18·25	—	—	—	—	—	—	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1
17·75	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
17·25	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1
16·75	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	1
16·25	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	—	—	—	—	—	—	—	1
15·75	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	1
15·25	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	—	—	1	—	—	—	—	2
14·75	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
14·25	—	—	—	—	—	—	—	—	—	—	—	—	1	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	2
13·75	—	—	—	—	1	—	—	—	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	1	—	—	—	2
13·25	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	—	—	2	—	—	—	—	—	—	4
12·75	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	—	—	—	—	1
12·25	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
11·75	—	—	—	—	—	1	1	1	—	—	1	1	—	—	—	—	0·5	0·5	2	1	—	—	—	—	—	—	—	—	—	5
11·25	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	0·5	0·5	1	—	0·5	0·5	—	—	—	—	—	—	—	—	5
10·75	—	—	—	—	—	—	—	—	—	—	—	1	1	—	1	—	1	—	—	—	—	—	—	—	—	—	—	—	—	5
10·25	—	—	—	—	—	—	—	—	1	1	—	1	1	—	—	1	—	—	—	—	—	1	—	—	—	—	—	—	—	5
9·75	—	—	—	—	1	—	—	1	1	—	—	1	1	2	—	—	—	—	1	1	—	—	—	—	—	—	—	—	—	9
9·25	—	—	—	—	—	—	—	—	1	—	1	1	1	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	6
8·75	1	1	—	—	—	—	1·25	1·25	—	1	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	5·5
8·25	—	—	—	—	—	—	0·25	1·25	—	1	1	—	—	—	2	—	—	—	—	—	—	—	—	—	—	—	—	—	—	5·5
7·75	—	—	—	1	—	—	—	1	1·5	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	4·5
7·25	—	—	1	1	—	—	—	2	1·5	—	1	1	2	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	9·5
6·75	—	—	1	—	2	1	2	2	—	1	4	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	14
6·25	—	—	1	—	2	1	2	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	7
5·75	1	—	1	1	—	3	—	1	1	1	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	10
5·25	—	2	—	1	3	1	4·25	3·25	2	2	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	18·5
4·75	—	—	—	3	—	2	1·75	0·75	3	1·5	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	12
4·25	—	—	—	3	2·5	5	4	5	0·5	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	20
3·75	—	2	1	3	4·5	2·5	1·5	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	14·5
3·25	—	1	2	1	5	4	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	13
2·75	—	2	2	1	3	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	9
2·25	—	1	—	3	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	5
1·75	—	1	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	2
1·25	1	2	2	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	5
0·75	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1
0·25	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Totals	3	12	11	19	25	21·5	18	20·5	12·5	8·5	9	7	8	5	6	1·5	1·5	3	4·5	1·5	2	1	2	2	1	1	—	—	1	208

The theoretical coefficient of correlation, the regression coefficients and their standard errors for 208 samples are compared with observations below:

	Theory	Standard Error	Observation
Correlation, $r_{\psi\psi_1}=\sqrt{\frac{k-c-1}{k-1}}=\sqrt{\frac{4}{7}}$	$=\cdot7559$	$\cdot0297$	$\cdot7850$
Regression of ψ on ψ_1	$=1\cdot0000$	$\cdot0600$	$1\cdot0219$
Regression of ψ_1 on $\psi=\frac{k-c-1}{k-1}$	$=\cdot5714$	$\cdot0343$	$\cdot6031$

The agreement is very satisfactory. The regression lines and observed array means are shown in Figures 4 and 5. The means in Figure 4 lie fairly systematically below the regression line, a divergence resulting from the fact that the observed mean ψ is 6·68 instead of 7·00.

FIG. 4



(*d*) *The Distribution in Arrays of constant* χ_1^2.

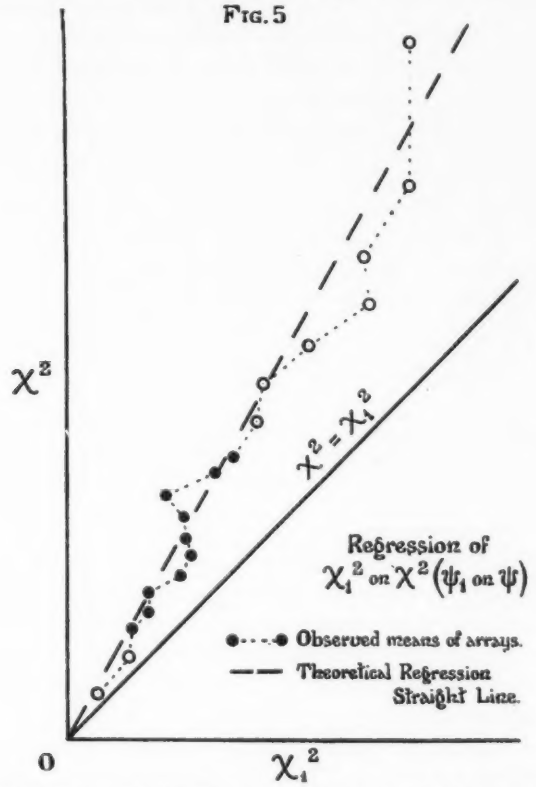
Writing $\theta=\chi^2-\chi_1^2=\psi-\psi_1$, the theoretical distribution in each array is a Type III curve with start on the line $\psi=\psi_1$, and equation $y=y_0\theta^{\frac{1}{2}}e^{-\frac{1}{2}\theta}$, for which

$$\begin{cases}\text{Mean } \theta=c=3\cdot0000 \text{ and standard error for 208 samples of } 0\cdot1698,\\ \sigma_\theta^2=2c=6\cdot0000 \quad\quad " \quad\quad " \quad\quad " \quad\quad 1\cdot0190^*.\end{cases}$$

* Using the expression for the standard error given in preceding footnote.

The frequencies in single arrays are not adequate to provide a test comparison, but as the arrays should be homoscedastic they may be combined together, and the observed distribution of the 208 values of θ compared with theory. It is found that

FIG. 5



Mean $\theta = 2{\cdot}6382$*, $\sigma_\theta^2 = 4{\cdot}9637$. The latter differs from the theoretical value by just over its standard error, but the former by more than twice that amount. Comparing frequencies as shown in Table II, it is found on applying the Test for

TABLE II.

Experimental and Theoretical Distribution of $\theta = \chi^2 - \chi_1^2$.

$\theta = \chi^2 - \chi_1^2$	0·0—1·0	1·0—2·0	2·0—3·0	3·0—4·0	4·0—5·0	5·0—6·0	> 6·0	Totals
Theory ...	41·3	47·6	37·6	27·1	18·7	12·5	23·2	208
Observation ...	53·0	48·0	35·0	28·5	16·0	11·5	16·0	208

* Owing to the groupings used, the values of Mean $\theta = 2{\cdot}6382$ and Mean ψ – Mean $\psi_1 = 2{\cdot}6356$, do not exactly agree.

Goodness of Fit that with the 7 groups $\chi^2 = 6{\cdot}25$ and $P = {\cdot}40$, and although this in itself suggests quite a good fit, it is clear that the observed values of θ are rather too small, with an excess in the lowest group and a defect in the highest.

The divergence of Mean ψ from theory of just over its standard error may very probably be due to sampling fluctuations, but owing to the high correlation of ψ with ψ_1 a low value in Mean ψ should be associated with a low value of Mean ψ_1. Actually it is not, and so the difference or Mean θ is unexpectedly small. The probable explanation of this lies in the fact that as the process of fitting has been by moments, the values of ψ_1 may sometimes differ considerably from *minimum* ψ_1, consequently the Mean ψ_1 is too large and Mean θ too small.

In order to see whether there is any tendency for the distributions in the arrays of constant ψ_1 to differ more widely from theory when ψ_1 is large, the mean and variance of θ were calculated for those arrays in which $\chi_1^2 = \psi_1 > 6{\cdot}00$. The 41 cases gave the following result:

	Theory	Standard Error	Observation
Mean	3·00	0·38	2·90
Variance	6·00	2·05	6·28

The observations therefore provide no evidence that the minimum-ψ_1 theory fails here, although disagreement might be expected to arise in this part of the table which corresponds to regions in the k-dimensioned space far removed from the centre of the density field.

It is clear, therefore, that on the whole these results provide a satisfactory confirmation of the theory, as far as this can be obtained from 208 samples. The labour involved in increasing their number appeared prohibitive. The main discrepancy lies in the difference between an observed Mean $(\chi^2 - \chi_1^2)$ of 2·64 and the theoretical value of 3·00, but, as has been pointed out, this is very probably due to the fact that we have not actually calculated the minimum χ_1^2.

(5) *Summary of the Position.*

In 1922 Mr G. U. Yule* described a series of sampling experiments from a population classified for two characters in the form of a contingency table in which there was no association, and showed that the observed distribution of χ^2 was of the standard form while that of χ_1^2 appeared to be adequately represented by the modified distribution given by R. A. Fisher. The sampling experiment of the preceding section confirms and supplements Yule's results in another problem. These results are of course only single instances, but they certainly suggest that the approximations involved in reaching not only the theoretical χ^2 and minimum χ_1^2 distributions but also the frequency relation between the two do not lead to forms which are merely limits for very large samples and out of touch with practical conditions. The value calculated for χ_1^2 does, however, depend upon the method of fitting employed, and though in a great many cases it may be easy to feel reasonably certain that the method of moments, or of least squares or other simple

* *Jour. Roy. Stat. Soc.* LXXXV. pp. 95—104.

solution provides a χ_1^2 approaching the minimum, there is a real difficulty in this point. It has always been realised, for example, that in dealing with very skew or abrupt distributions the use of uncorrected moments is an inefficient method of fitting, but the process of correction does not appear to be a simple one in practice, particularly when as is often the case the frequency groups in which the data are presented to the statistician are most inconveniently arranged.

Let us suppose however that a satisfactory method of fitting has been employed, and that χ_1^2 is approaching the minimum; we may return to the second form of question put on p. 268 above: *Is it probable that the observed sample would have arisen in random sampling from a population whose law of frequency is defined by an equation depending on c undetermined parameters?* This question is of a different form from that which arises in the standard χ^2 problem where a population is exactly specified. The mere fact that χ_1^2 is a quantity easily calculated and one whose frequency distribution in sampling is under certain conditions known, is not in itself an adequate reason for making use of χ_1^2 as a criterion and basing one's judgment on the value of P, the probability integral of the distribution. The P calculated from the ordinary χ^2 measures the probability of drawing samples from a given population for which the chance of occurrence is equal or less than that of the observed sample; the contours of χ are in fact contours of equal density in the k-dimensioned space. In the present case, however, we are not dealing with a hypothesis regarding a single population but regarding a population law—in fact, regarding the population locus in the k-dimensioned space instead of a single population point. This is what we have termed a composite hypothesis. By supposing a certain *à priori* probability distribution of points on the population locus and superposing the corresponding density fields, it might be possible to show that the contours of constant minimum χ_1^2 were approximately the contours of equal probability *à posteriori* and so bring the methods of solution of the two problems into agreement. But it would seem that the use of the criterion of likelihood leads in a simple way to complete consistency between the two solutions.

With the notation of our definition of likelihood, Ω is the set of populations with all possible values of the group proportions $p_s, (s = 1, 2 \ldots k)$, while the composite hypothesis to be tested is that Π belongs to the sub-set ω of populations satisfying the equations (16). The likelihood of this hypothesis is given by the ratio λ_1, which just as the λ of (12) will be

$$\lambda_1 = \frac{C(\omega_{\max.})}{C(\Omega_{\max.})} = \left(\frac{m_1}{n_1}\right)^{n_1} \ldots \left(\frac{m_k}{n_k}\right)^{n_k} \quad \ldots\ldots\ldots\ldots\ldots\ldots (43),$$

where now we give to $m_s = Np_s$ those values satisfying the equations (16) which make λ_1 a maximum. But we have seen approximately $\lambda_1 = e^{-\frac{1}{2}\chi_1^2}$, and consequently the values of m_s which make λ_1 a maximum will be approximately those making χ_1 a minimum. It follows that the contours of constant likelihood for the composite hypothesis are approximately the contours of constant minimum χ_1 or the envelopes of the hyperellipsoids. We should, therefore, base our judgment regarding the probability of the hypothesis on the probability integral of χ_1, and

this as we have seen should take approximately the same form from whatever population divided into k groups we sample, namely,

$$P_1 = \frac{\int_{\chi_1}^{\infty} \chi^{k-c-2} e^{-\frac{1}{2}\chi^2} d\chi}{\int_{0}^{\infty} \chi^{k-c-2} e^{-\frac{1}{2}\chi^2} d\chi} \quad \text{.....(44).}$$

It is of interest to note that we have here a further illustration of a general proposition. *The contours of constant likelihood for a multiple hypothesis are the envelopes of the corresponding contours associated with the constituent simple hypotheses.* In our previous paper, when dealing with samples from normal populations, we described the simple hypothesis as Hypothesis A; for this the contours of likelihood in the fundamental N-dimensioned space were certain figures of revolution. Hypothesis B was defined as the multiple hypothesis concerning the sub-set of populations with a fixed mean but varying standard deviation; the contours of likelihood were in this case the hypercones on which Student's ratio z was constant. These hypercones were the envelopes of the contours of Hypothesis A. The integral of the density outside or between these envelopes was completely independent of the particular population with which the density was associated. A similar result holds in the (χ^2, χ_1^2) problem, although here only as an approximation. For the test to be possible this independence between the density-integral outside a given envelope and the population point, with which the density is associated, is essential. How far this rule will hold in general we do not know.

Now whether the principle of likelihood be accepted or no, it is at once clear that the calculation of the P_1 of (44) provides a method of comparing from one point of view the relative merits of two different frequency laws, applied to a single sample. If a normal curve and a 6th order parabola give both a value of $\chi_1^2 = 12{\cdot}0$ when fitted to a sample divided into 15 frequency groups, we can hardly avoid an instinctive feeling that the former curve provides the more satisfactory representation of the data. And this is not from some general prejudice in favour of a normal curve. We feel that it is a case for the use of "Occam's Razor," where the simpler hypothesis should be adopted. On entering the Tables with $\chi^2 = 12{\cdot}0$ and $n' = 15$ and finding $P = {\cdot}61$ in both cases, we obtain no numerical support for this common-sense feeling. But, on the other hand, if for the first curve we take $n' = k - c = 15 - 2 = 13$, and for the second $n' = 15 - 6 = 9$, we find the two values for P_1 of ·45 and ·15, and can express what we feel intuitively in more exact form.

There is an alternative method of approaching this problem which makes use of the (χ^2, χ_1^2) frequency relation to predict χ^2 from χ_1^2. The reasoning would be as follows. If the sample had in fact been drawn from a population following the supposed law, the average value of χ^2 in repeated samples associated with the observed χ_1^2 is the mean value of χ^2 in the corresponding array of χ_1^2, or $\chi_1^2 + c$ (see p. 273, para. (*c*)). Taking this value we should enter the Tables with $\chi^2 = \chi_1^2 + c$ and $n' = k$, thus attempting to apply the ordinary χ^2 interpretation to a corrected χ_1^2.

The results of these methods may be compared and summarised in an example. Suppose that we have a sample divided into 16 groups and that we wish to judge whether it is likely to have come from a population following a Type IV distribution law. We obtain the "best fitting" curve and find $\chi_1^2 = 19{\cdot}00$. Here $c = 4$ and we may proceed as follows:

(*a*) $\chi^2 = \chi_1^2 = 19{\cdot}00$, $n' = 16$, $P = {\cdot}214$.

(*b*) $\chi^2 = \chi_1^2 = 19{\cdot}00$, $n' = 16 - 4 = 12$, $P_1 = {\cdot}061$.

(*c*) $\chi^2 = \chi_1^2 + c = 23{\cdot}00$, $n' = 16$, $P_2 = {\cdot}084$.

In the first case we argue that if the population had corresponded to the fitted curve or very nearly to this, the sample would be quite a reasonable one, but as we realise that we are dealing with a best-fitting curve we decide not to lay too much stress on the goodness of fit. In case (*b*) we see that the criterion is low, but that in 6 per cent. of cases in sampling from a Type IV curve it would be as small or smaller; we again hesitate to draw any definite conclusion. In the last case we attempt to judge in another way (avoiding any appeal to the criterion of likelihood) the effect of having used 4 constants in the fitted curve.

The three different values of P are not in fact inconsistent, because the interpretation of each should be coordinated with a different process of reasoning. To some extent they supplement one another, and it is suggested that in drawing conclusions from the fitting of a curve it will often be of value to obtain the P of both (*a*) and (*b*); but beyond this we do not put forward one method as right and another as wrong, for it is far better that each reader should grapple the problem for himself, and having understood the position, make use of the test which seems most relevant to him, indicating at the same time upon what basis he is arguing.

(6) *The Case of Two Samples.*

This problem may be taken as an illustration of the result of the preceding section. Suppose that we have two samples of size N and N', for each of which the observations are classed into t groups in which the frequencies are

$$n_1, n_2 \ldots n_t \text{ and } n_1', n_2' \ldots n_t'.$$

We wish to test the hypothesis that these two samples have been drawn from the same population, although we are not interested in its exact form. In applying the method of likelihood we shall take as Ω the set of all possible pairs of populations with group proportions p_s and p_s' respectively ($s = 1, 2 \ldots t$). The chance of drawing the observed pair of samples is then

$$C = \frac{N!\,N'!}{n_1! \ldots n_t!\, n_1'! \ldots n_t'!} p_1^{n_1} \ldots p_t^{n_t} p_1'^{n_1'} \ldots p_t'^{n_t'} \quad \ldots\ldots\ldots\ldots (45),$$

which we know from an extension of (4) and (10) is approximately equal to

$$C = C_0 e^{-\frac{1}{2}\chi^2} \quad \ldots\ldots\ldots\ldots (46),$$

where C_0 is a constant independent of the n_s and n_s' but depending on the p_s and p_s', while

$$\chi^2 = \sum_{s=1}^{t} \left\{ \frac{(n_s - Np_s)^2}{Np_s} + \frac{(n_s' - N'p_s')^2}{N'p_s'} \right\} \quad \ldots\ldots\ldots\ldots (47).$$

Since $n_1 + \ldots + n_t = N$ and $n_1' + \ldots + n_t' = N'$, the point representing the pair of samples can be represented in a field of $2t-2$ dimensions with hyperellipsoidal density contours.

We wish now to test the likelihood of the composite hypothesis that the two sampled populations are identical, that is to say, that they belong to the sub-set ω of the set Ω in which $p_s = p_s' = q_s$, say, for $s = 1, 2 \ldots t-1$. The point representing the populations of origin in the $(2t-2)$-dimensioned space must therefore lie on $t-1$ primes defined by

$$p_s = p_s' \text{ or } m_s N' = m_s' N \qquad (s = 1, 2 \ldots t-1).$$

As in the general case we must therefore calculate in any given problem the minimum value of

$$\chi_1^2 = \sum_{s=1}^{t} \left\{ \frac{(n_s - Nq_s)^2}{Nq_s} + \frac{(n_s' - N'q_s)^2}{N'q_s} \right\} \quad \ldots\ldots\ldots\ldots\ldots\ldots (48),$$

and use this value of χ_1 in the expression (44) for P_1. But we shall now have

(a) the $2t-2$ dimensions corresponding to the $k-1$ of the general case,

(b) the $t-1$ primes of the population locus corresponding to the $k-c-1$.

Hence we have

$$P_1 = \frac{\int_{\chi_1}^{\infty} \chi^{t-2} e^{-\frac{1}{2}\chi^2} d\chi}{\int_{0}^{\infty} \chi^{t-2} e^{-\frac{1}{2}\chi^2} d\chi} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (49).$$

But the minimum value of χ_1^2 of (48) occurs when $q_s = (n_s + n_s')/(N + N')$, and using these values it is found at once that

$$\chi_1^2 = \sum_{s=1}^{t} \left\{ \frac{NN' \left(\frac{n_s}{N} - \frac{n_s'}{N'} \right)^2}{n_s + n_s'} \right\} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (50).$$

The test involving the use of (49) and (50) corresponds exactly to that given by K. Pearson* for testing the significance of the difference between two grouped samples.

(7) *Application to Contingency Tables.*

This was the original problem considered by Fisher and Yule in the papers referred to above†. If the observations of the sample are divided into $p \times q$ categories which can be arranged in the form of a contingency table, the problem of testing whether ϕ^2, the coefficient of mean square contingency, differs significantly from zero is the same as that of asking whether it is likely that the sample has been drawn from a population whose distribution law is what may be termed an "independence bivariate histogram" or more simply an independence distribution. Such a population law is defined by $p+q$ positive marginal ratios $a_1, a_2 \ldots a_p$; $b_1, b_2 \ldots b_q$, which are restricted by the relations

$$a_1 + a_2 + \ldots + a_p = 1, \quad b_1 + b_2 + \ldots + b_q = 1 \quad \ldots\ldots\ldots\ldots\ldots (51).$$

* *Biometrika*, VIII. pp. 250—254.

† *Jour. Roy. Stat. Soc.* Vol. LXXXV. pp. 87—94 and 95—104.

The expected sample frequencies in the (s, t)th cell are then

$$m_{st} = N a_s b_t.$$

If we adopt the usual notation and write $n_{s.}$ and $n_{.t}$ for the marginal totals of the sth column and tth rows in the sample, we shall have for a known population

$$\chi^2 = S_{s,t} \frac{(n_{st} - \widetilde{m}_{st})^2}{\widetilde{m}_{st}} \quad \text{.....(52),}$$

and for an independence distribution fitted to the sample,

$$\chi_1^2 = N\phi^2 = S_{s,t} \frac{(n_{st} - m_{st})^2}{m_{st}} = S_{s,t} \frac{\left(n_{st} - \frac{n_{s.} n_{.t}}{N}\right)^2}{\frac{n_{s.} n_{.t}}{N}} \quad \text{.....(53).}$$

The values $m_{st} = n_{s.} n_{.t}/N$ used in (53) are not those which make χ_1^2 a minimum, but are those making the likelihood that the sample comes from an independence population a maximum.

For
$$\lambda \propto \frac{N!}{n_{11}! n_{12}! \dots n_{pq}!} (a_1 b_1)^{n_{11}} (a_1 b_2)^{n_{12}} \dots (a_p b_q)^{n_{pq}}$$
$$= \frac{N!}{n_{11}! \dots n_{pq}!} a_1^{n_{1.}} a_2^{n_{2.}} \dots a_p^{n_{p.}} b_1^{n_{.1}} b_2^{n_{.2}} \dots b_q^{n_{.q}} \quad \text{.....(54),}$$

and for a given sample

$$\log \lambda = \text{constant} + n_{1.} \log a_1 + \dots + n_{p.} \log a_p + n_{.1} \log b_1 + \dots + n_{.q} \log b_q.$$

Hence differentiating, equating to zero and remembering that only $p-1$ of the a's and $q-1$ of the b's are independent, it is found that

$$\left.\begin{aligned} n_{1.}/a_1 = n_{2.}/a_2 = \dots = n_{p.}/a_p = N \\ n_{.1}/b_1 = n_{.2}/b_2 = \dots = n_{.q}/b_p = N \end{aligned}\right\} \quad \text{.....(55),}$$

or the solution of maximum likelihood is

$$a_s = n_{s.}/N, \quad b_t = n_{.t}/N, \quad m_{st} = N a_s b_t = n_{s.} n_{.t}/N.$$

Fisher has shown in the general case that for large samples the solution of maximum likelihood tends to that which minimises χ_1^2*, and if therefore it may be assumed that the sampling distribution of the χ_1^2 of (53) is approximately that of a minimum χ_1^2, the approximate distribution of ϕ^2 in sampling from an independence population follows at once. The independence distribution fitted to the sample contains $p+q-2$ independent constants, and therefore writing in (39) $k = pq$, $c = p+q-2$ and $\psi_1 = N\phi^2$, we have for the distribution of ϕ^2 the Type III curve,

$$y = y_0' (\phi^2)^{\frac{(p-1)(q-1)}{2} - 1} e^{-\frac{1}{2}N\phi^2} \quad \text{.....(56),}$$

which has Mean $\phi^2 = (p-1)(q-1)/N$, $\sigma_{\phi^2} = \sqrt{2(p-1)(q-1)}/N$.

This result can, however, be obtained by a method which makes no appeal to the conceptions of maximum likelihood and minimum χ_1^2. The set of all possible samples of N from a fixed independence population with marginal ratios $\tilde{a}_s$, $\tilde{b}_t$ $(s = 1, 2 \dots p;\ t = 1, 2 \dots q)$ can be divided into a number of sub-sets for each of

* *Phil. Trans.* Vol. 222 A, p. 357; *Jour. Roy. Stat. Soc.* Vol. LXXXVII. p. 447.

which the marginal totals of the sample have the same series of values. Consider the frequency distribution among the cells in the samples composing one of these sub-sets in which, let us say, the marginal totals are $\nu_{s.}$, $\nu_{.t}$ ($s = 1, 2 \dots p$; $t = 1, 2 \dots q$). The frequency of occurrence of a particular combination of cell contents n_{st} will be proportional to a term of the multinomial expansion, or as in (54) to

$$\frac{N!}{n_{11}!\, n_{12}! \dots n_{pq}!} \tilde{a}^{\nu_{1.}} \dots \tilde{a}_p^{\nu_{p.}} . \tilde{b}_1^{\nu_{.1}} \dots \tilde{b}_q^{\nu_{.q}},$$

or since the ν's are constant within the sub-set, to $N!/(n_{11}!\, n_{12}! \dots n_{pq}!)$. That is to say, the distribution of frequency and hence the distribution of $\chi_1^2 = N\phi^2$ within the sub-set does not depend upon the $\tilde{a}$'s and $\tilde{b}$'s of the sampled population, but only on the specially chosen marginal totals $\nu_{s.}$ and $\nu_{.t}$. Within this sub-set, therefore, defined by the ν's the distribution of cell contents in random sampling will be the same whatever be the independence distribution sampled; in fact we may suppose that its marginal ratios actually have the values $\tilde{a}_s = \nu_{s.}/N$, $\tilde{b}_t = \nu_{.t}/N$ given by the marginal totals of the sub-set. But if this were so, the distribution of χ_1 within the sub-set becomes simply that of a χ obtained for samples of N for which the frequencies are restricted by $p + q - 2$ linear relations,

$$\left.\begin{aligned} n_{s1} + n_{s2} + \dots + n_{sq} &= \nu_{s.} \qquad (s = 1, 2 \dots p-1) \\ n_{1t} + n_{2t} + \dots + n_{pt} &= \nu_{.t} \qquad (t = 1, 2 \dots q-1) \end{aligned}\right\} \quad \dots\dots\dots\dots(57).$$

We are now dealing no longer with envelopes but with hyperellipsoids in a space of reduced dimensions, and for this case it is known as proved originally in *Biometrika*, XI.* that the distribution of χ is given by

$$y = y_0 \chi^{pq-2-(p+q-2)} e^{-\frac{1}{2}\chi^2} \quad \dots\dots\dots\dots(58).$$

This distribution for χ_1 will hold within each of the sub-sets†; it will therefore hold for the aggregate of all possible samples seeing that it is independent of the ν's defining the sub-set. It follows that the distribution of $\phi^2 = \chi_1^2/N$ is that given in (56). The possibility of this method of solution appears to arise from the special form of the population locus representing independence distributions, and the method does not seem to be rigorously applicable in general.

It is perhaps of interest to conclude by referring to a third and entirely different way in which one of us approached the ϕ^2 distribution a few years ago.

Writing $n_s = \tilde{m}_{st} + \delta n_{st}$, $n_{s.} = \tilde{m}_{s.} + \delta n_{s.}$, $n_{.t} = \tilde{m}_{.t} + \delta n_{.t}$, it is found that

$$\chi_1^2 = N\phi^2 = S_{st}\left\{\frac{(\tilde{m}_{st} + \delta n_{st} - (\tilde{m}_{s.} + \delta n_{s.})\,(\tilde{m}_{.t} + \delta n_{.t})/N)^2}{(\tilde{m}_{s.} + \delta n_{s.})\,(\tilde{m}_{.t} + \delta n_{.t})/N}\right\}$$

$$= S_{st}\left\{\frac{\delta n_{st}^2}{\tilde{m}_{st}} + \frac{\tilde{m}_{st}}{\tilde{m}_{s.}^2}\delta n_{s.}^2 + \frac{\tilde{m}_{st}}{\tilde{m}_{.t}^2}\delta n_{.t}^2 - \frac{2\delta n_{st}\delta n_{s.}}{\tilde{m}_{s.}} - \frac{2\delta n_{st}\delta n_{.t}}{\tilde{m}_{.t}} + \frac{2\delta n_{s.}\delta n_{.t}}{N}\right\} \dots(59),$$

* K. Pearson, "On the Theories of Multiple and Partial Contingency," p. 145.

† There is of course an approximation involved at this point, because even if the sample be very large, in *some* of the sub-sets the frequencies $m_{st} = \nu_{s.}\nu_{.t}/N$ must still be too small to justify the assumptions leading to the χ^2 distribution. The chance of a sample falling into one of these sub-sets is, however, so small as barely to affect the form of the aggregate distribution of all possible samples.

going only as far as second order terms. The mean value of the squares and products of the cell variations in (59) are well known, and it was easy to show on taking mean values that

$$\text{Mean } \chi_1^2 = (p-1)(q-1).$$

Next taking the mean value of the square of (59) and evaluating at some length the 4th order products of form $(\delta n_{st}\delta n_{hk}\delta n_{s.}\delta n_{h.})$, etc., it was possible to show that to the same order of approximation

$$\sigma_{\chi_1^2} = \sqrt{2(p-1)(q-1)}.$$

Further, since

$$\chi^2 = \sum_{st}\{\delta n_{st}^2/\tilde{m}_{st}\} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (60),$$

it can be shown in the same way that

$$\text{Mean } \chi^2 = pq-1, \quad \sigma_{\chi^2} = \sqrt{2(pq-1)}.$$

Multiplying together the expressions (59) and (60), taking mean values and evaluating the resulting 4th order products, it was found that

$$\text{Mean } (\chi^2\chi_1^2) = p^2q^2 + 2pq - p^2q - pq^2 - p - q + 1,$$

and from these results the coefficient of correlation between χ^2 and χ_1^2 is found to be

$$r_{\chi^2\chi_1^2} = \sqrt{\frac{(p-1)(q-1)}{pq-1}}.$$

Remembering that in the present case $k = pq$ and $c = p+q-2$, it will be seen that these values for the mean and standard deviation of χ^2 and χ_1^2 and for the correlation between the two correspond exactly to special cases of the general results given on p. 273 above. This is an interesting confirmation of the previous work and suggests that the order of approximation involved is in both cases the same.

APPENDIX.

The Distribution of χ_1.

In section (3) above the distribution of minimum χ_1 given in equation (36) was deduced by the use of geometrical arguments in multiple space, applied to a position which had been simplified by certain assumptions. We shall first give in the following Lemma what may appear a more exact proof to replace the geometrical reasoning of pp. 271—273 above, and then establish more rigorousiy the manner in which the true distribution of χ_1 tends to that which has been obtained from these simplifying assumptions.

Instead of the sample frequencies $n_1, n_2, \ldots n_k$, take as variables the proportions $q_1, q_2, \ldots q_k$ where $Nq_s = n_s$, and let $p_1, p_2, \ldots p_k$ be the corresponding population group proportions, where $Np_s = \tilde{m}_s$. It follows that

$$q_1 + q_2 + \ldots + q_k = 1, \quad p_1 + p_2 + \ldots + p_k = 1.$$

Let P_{χ_1} be defined by

$$P_{\chi_1} = c_1 \iint\ldots\int_{\mathcal{E}} e^{-\frac{N}{2}\sum_{s=1}^{k}\frac{(p_s-q_s)^2}{p_s}}\, dq_1 \ldots dq_{k-1} \quad \ldots\ldots\ldots\ldots\ldots\ldots (61),$$

where the region of integration is that *inside* the envelope $\mathcal{E}$, of the hypersurfaces

$$N\sum_{s=1}^{k}\left\{\frac{(p_s + \sum_{t=1}^{c}\alpha_{st}(\alpha_t - \alpha_t^{(0)}) - q_s)^2}{p_s}\right\} = \chi_1^2 = \text{constant} \quad \ldots\ldots\ldots\ldots (62),$$

the envelope being taken with regard to the $c < k-1$ parameters $\alpha_1, \alpha_2, \ldots \alpha_c$. The constant c_1 in (61) is such that $P_{\chi_1} \to 1$ as $\chi_1 \to \infty$.

Lemma.

If the number of parameters, α, specifying the hypersurfaces cannot be reduced, then

$$P_{\chi_1}=\int_0^{\chi_1}\chi^{k-c-2}e^{-\frac{1}{2}\chi^2}d\chi\Big/\int_0^{\infty}\chi^{k-c-2}e^{-\frac{1}{2}\chi^2}d\chi \quad \text{.....(63).}$$

To prove this Lemma we shall show that P_{χ_1} is proportional to an integral of the form

$$\iint\dots\int_H e^{-\frac{1}{2}\sum_{s=c+1}^{k-1}(A_s x_s^2)}\,dx_{c+1}\dots dx_{k-1} \quad \text{.....(64),}$$

the region of integration H being that inside the hypersurface

$$\sum_{s=c+1}^{k-1}\{A_s x_s^2\}=\chi_1^2 \quad \text{.....(65).}$$

The condition that the number of parameters, α, specifying the hypersurfaces (62) cannot be reduced, means that at least one of the determinants of cth order of the matrix

$$\begin{vmatrix} a_{11} & a_{12}\dots a_{1c}\\ a_{21} & a_{22}\dots a_{2c}\\ \dots\dots\dots\dots\\ \dots\dots\dots\dots\\ a_{k1} & a_{k2}\dots a_{kc}\end{vmatrix}$$

is not zero. We may always assume that

$$\begin{vmatrix} a_{11} & a_{12}\dots a_{1c}\\ a_{21} & a_{22}\dots a_{2c}\\ \dots\dots\dots\dots\\ \dots\dots\dots\dots\\ a_{c1} & a_{c2}\dots a_{cc}\end{vmatrix}\neq 0,$$

and thus that the system of c linear equations

$$q_s'=p_s+\sum_{t=1}^{c}a_{st}(\alpha_t-\alpha_t^0) \quad (s=1, 2, \dots c) \quad \text{.....(66)}$$

can be solved with regard to $(\alpha_t-\alpha_t^0)$ giving solutions which are not all equal to zero.

Clearly the hypersurfaces (62) considered in the $(k-1)$-dimensioned space of $q_1, q_2, \dots q_{k-1}$ are hyperellipsoids with centres at the variable point,

$$q_s'=p_s+\sum_{t=1}^{c}a_{st}(\alpha_t-\alpha_t^0) \quad (s=1, 2, \dots k-1) \quad \text{.....(67).}$$

These centre points must lie on the intersection of $k-1-c$ primes, say Q. In fact, solving the equations (66) with regard to $(\alpha_t-\alpha_t^0)$ and substituting the results into the last $k-1-c$ equations of (67), we obtain this number of first order equations among $q_1', q_2', \dots q_{k-1}'$ which will be satisfied by every centre of the hyperellipsoids (62).

We shall now change the system of coordinates, taking as origin the point $p_1, p_2, \dots p_{k-1}$, and for coordinate primes

(a) the $k-1-c$ primes Q,

(b) c primes with directions as in the previous system, choosing these arbitrarily except for the condition that none of them are parallel to the members of Q.

If the equations of the primes Q be

$$\sum_{s=1}^{k-1}b_{sj}(q_s-p_s)=0 \quad (j=1, 2, \dots k-1-c) \quad \text{.....(68)}$$

then the formulae of transformation will be

$$\left.\begin{aligned} x_j&=A_j\sum_{s=1}^{k-1}b_{sj}(q_s-p_s) && (j=1, 2, \dots k-1-c) \quad \text{.....(69)}\\ x_l&=A_l(q_l-p_l) && (l=k-c, k-c+1, \dots k-1) \quad \text{.....(70)}\end{aligned}\right\}$$

where the A's are some constants not all equal to zero. (The determinant of the transformation is not zero, since all the coordinate primes cut one another.) The new coordinates of the centres of the hyperellipsoids (62), say β's, will be $\beta_j=0$, $(j=1,2,\ldots k-1-c)$ and $\beta_l=f_l(\alpha_1,\ldots\alpha_c)$ $(l=k-c,\ldots k-1)$ where the f's are some linear functions of the α's. We can now take the β's instead of the α's as the c variable parameters specifying the hyperellipsoids (62). The new equations of these hyperellipsoids will now take the form

$$\sum_{s=1}^{k-1}\sum_{t=1}^{k-1}\{B_{st}(x_s-\beta_s)(x_t-\beta_t)\}=\chi_1^2 \quad\ldots\ldots(71),$$

where the B's are constants satisfying the conditions $B_{st}=B_{ts}$ and $B_{ss}>0$, and where we must remember that $\beta_s=0$ for $s=1, 2, \ldots k-1-c$.

By collecting together all the terms in (71) containing the factor $(x_{k-1}-\beta_{k-1})$, and adjusting to form a perfect square, we may put this expression in the form

$$\left[\sum_{s=1}^{k-1}\{\gamma_{k-1,s}(x_s-\beta_s)\}\right]^2+\sum_{s=1}^{k-2}\sum_{t=1}^{k-2}\{B_{st}'(x_s-\beta_s)(x_t-\beta_t)\}=\chi_1^2 \quad\ldots\ldots(72),$$

and repeating this operation to the second term on the left-hand side of (72), and successively to the product terms obtain finally the following form of equation,

$$\sum_{i=1}^{k-1}\{\phi_i^2\}=\chi_1^2 \quad\ldots\ldots(73),$$

where ϕ_i stands for $\sum_{s=1}^{i}\{\gamma_{is}(x_s-\beta_s)\}$, the γ's being constants and $\gamma_{ii}\neq0$*.

We may now calculate the envelope $\mathcal{E}'$ of the transformed hyperellipsoids, using for that purpose their equations in the form (73). Differentiating with regard to the c parameters $\beta_{k-1}, \beta_{k-2}, \ldots \beta_{k-c}$ and equating the derivatives to zero, we find,

$$\left.\begin{aligned}
-\frac{\partial\chi_1^2}{\partial\beta_{k-1}}&=2\gamma_{k-1,k-1}\phi_{k-1}=0\\
-\frac{\partial\chi_1^2}{\partial\beta_{k-2}}&=2\gamma_{k-2,k-2}\phi_{k-2}+2\gamma_{k-1,k-2}\phi_{k-1}=0\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
-\frac{\partial\chi_1^2}{\partial\beta_{k-c}}&=2\sum_{s=k-c}^{k-1}\{\gamma_{s,k-c}\phi_s\}=0
\end{aligned}\right\}\quad\ldots\ldots(74).$$

* In the $(k-1)$-dimensioned space of q's the left-hand side of (62) is positive at every point except $q_s=p_s+\sum_{t=1}^{c}a_{st}(a_t-a_t^0)$ $(s=1, 2, \ldots k-1)$ where it is zero. It follows that in the transformed space of x's the left-hand side of (71) must be positive except at the corresponding point $\beta_1, \beta_2, \ldots \beta_{k-1}$, where it is zero. The first step in the process of rearrangement of (71) into (73) gives

$$\chi_1^2=\left\{\sqrt{B_{k-1,k-1}}(x_{k-1}-\beta_{k-1})+\frac{B_{k-1,k-2}}{\sqrt{B_{k-1,k-1}}}(x_{k-2}-\beta_{k-2})+\ldots\right\}^2+\left(B_{k-2,k-2}-\frac{B_{k-1,k-2}^2}{B_{k-1,k-1}}\right)(x_{k-2}-\beta_{k-2})^2$$
$$+\text{squared terms containing } (x_t-\beta_t)^2, \quad (t=k-3,\ldots 2, 1)$$
$$+\text{product terms containing } (x_t-\beta_t), \quad (t=k-2, k-3,\ldots 2, 1).$$

This expression must be *positive* at every point on the straight line defined by

$$x_t=\beta_t\,(t=k-3,\ldots 2, 1);\ B_{k-1,k-1}(x_{k-1}-\beta_{k-1})+B_{k-1,k-2}(x_{k-2}-\beta_{k-2})=0$$

except at $\beta_1, \beta_2, \ldots \beta_{k-1}$ where it vanishes. Hence $B_{k-1,k-1}B_{k-2,k-2}-B_{k-1,k-2}^2>0$ and $\gamma_{k-2,k-2}\neq0$; the process of extracting positive squared terms may thus be continued to the next stage. It may be shown in the same way that it can be continued step by step until (73) is reached, and that always $\gamma_{ii}\neq0$.

It follows from (74) that $\phi_s=0$, for $s=k-c, k-c+1, \dots k-1$. We need not, however, solve these equations for the β's, since clearly the result of the substitution of their solutions in (73) will be to give

$$\sum_{i=1}^{k-c-1}\{\phi_{io}^2\}=\chi_1^2 \quad \dots\dots(75),$$

where by ϕ_{io} is denoted the expression for ϕ_i in which we put $\beta_s=0$, or $\phi_{io}=\sum_{s=1}^{i}\{\gamma_{is}x_s\}$. (75) is the transformed equation of the envelope, $\mathcal{E}'$.

We must now transform the function under the integral in (61). $N\sum_{s=1}^{k}\frac{(p_s-q_s)^2}{p_s}=\chi^2$ represents in the original coordinate space a hyperellipsoid centred at $p_1, p_2, \dots p_{k-1}$. After the transformation it becomes that member of the set (73) which has its centre at the origin, that is for which $\beta_s=0$, $s=1, 2, \dots k-1$. Hence (61) transforms into

$$P_{\chi_1}=c_2\iint\dots\int_{\mathcal{E}'}e^{-\frac{1}{2}\sum_{i=1}^{k-1}\{\phi_{io}^2\}}\,dx_1dx_2\dots dx_{k-1} \quad \dots\dots(76).$$

The proof of the Lemma now follows at once. Since ϕ_{io} for $i=1, 2, \dots k-c-1$ is independent of $x_{k-c}, x_{k-c+1}, \dots x_{k-1}$, the left-hand side of the expression (75) is independent of these variables. Consequently in integrating inside $\mathcal{E}'$ we have no limitations for $x_{k-c}, \dots x_{k-1}$, and may write

$$P_{\chi_1}=c_2\times I\times\iint\dots\int_{\mathcal{E}'}e^{-\frac{1}{2}\sum_{i=1}^{k-c-1}\{\phi_{io}^2\}}\,dx_1\dots dx_{k-c-1} \quad \dots\dots(77),$$

where $$I=\int_{-\infty}^{+\infty}\dots\int_{-\infty}^{+\infty}e^{-\frac{1}{2}\sum_{i=k-c}^{k-1}\{\phi_{io}^2\}}\,dx_{k-c}\dots dx_{k-1}$$

and is independent of the values of $x_1, x_2, \dots x_{k-c-1}$.

But the integral in (77) is known to equal

$$\text{constant}\times\int_0^{\chi_1}\chi^{k-c-2}e^{-\frac{1}{2}\chi^2}d\chi$$

and this proves the Lemma, giving the expression (63) for the value of P_{χ_1}. This is the result previously obtained by the use of geometrical arguments.

We shall now consider the more general case. The process of fitting a theoretical law to the sample gives us a series of frequencies $m_s=N\pi_s$ $(s=1, 2, \dots k)$, where we shall suppose

$$\pi_s=p_s+\sum_{t=1}^{c}\{\alpha_{st}(a_t-a_t^0)\}+\phi_s(a_1, a_2, \dots a_c) \quad \dots\dots(78),$$

and the true value of χ_1^2 is

$$N\sum_{s=1}^{k}\left\{\frac{(\pi_s-q_s)^2}{\pi_s}\right\}=\chi_1^2 \quad \dots\dots(79).$$

If this expression represents minimum χ_1^2, then the integral we require is

$$z=C\iint\dots\int_E e^{-\frac{1}{2}\chi^2}dq_1\dots dq_{k-1} \quad \dots\dots(80)$$

where the region of integration is that inside the envelope, E, of the hypersurfaces (79), and χ is given by

$$\chi^2=N\sum_{s=1}^{k}\frac{(p_s-q_s)^2}{p_s} \quad \dots\dots(81),$$

while C is a constant such that $z\to 1$ as $\chi_1\to\infty$.

We shall now prove the following Theorem:

If the π's of (78) *are such that* (1) $\pi_1+\pi_2+\ldots+\pi_k=1$, (2) $\sum_{s=1}^{k}(\pi_s-p_s)^2=\rho^2$ *where*

$$\operatorname*{Lt}_{\rho\to 0}\left\{\sum_{s=1}^{k}(\phi_s^2/\rho^2)\right\}=0,$$

then as $N\to\infty$, Lt $(z)=P_{\chi_1}$ *of* (63).

The condition (2) implies that when the parameters α of the fitted curve tend to the population values α^0, the functions ϕ are indefinitely small quantities of an order higher than ρ.

Let $\Omega(\chi)$ be the region in the $(k-1)$-dimensioned space of q's inside the hyperellipsoid (81), and let

$$z_{\Omega(\chi)}=c\int\int\ldots\int_{\Omega(\chi)}e^{-\frac{1}{2}\chi^2}dq_1\ldots dq_{k-1} \quad\ldots\ldots(82).$$

Choose now a value χ_0 so large that the χ_1 of (79) is $<\chi_0$ and also

$$0\leqslant 1-z_{\Omega(\chi_0)}<\epsilon \quad\ldots\ldots(83),$$

where ϵ is an arbitrarily small positive quantity.

The method of proof will be to obtain two envelopes of the kind $\mathcal{E}$*, say $\mathcal{E}'$ and $\mathcal{E}''$, corresponding to different values of χ_1 which are such that at any rate within the region $\Omega(\chi_0)$, $\mathcal{E}'$ lies completely outside and $\mathcal{E}''$ completely inside the envelope E. We shall then show that the integrals corresponding to these envelopes, say $P_{\chi_1'}$ and $P_{\chi_1''}$, which satisfy the conditions

$$P_{\chi_1''}<z-\epsilon<P_{\chi_1'},$$

will both tend to P_{χ_1} as $N\to\infty$. With this object we shall first consider three hypersurfaces associated with a single point $(\pi_1, \pi_2, \ldots \pi_{k-1})$, i.e. with a single set of values of the α's, namely S_{r_1} and S_{r_2} of form

$$N\sum_{s=1}^{k}\frac{(\pi_s-\phi_s-q_s)^2}{p_s}=r^2 \quad\ldots\ldots(84),$$

and W of form

$$N\sum_{s=1}^{k}\frac{(\pi_s-q_s)^2}{\pi_s}=\chi_1^2 \quad\ldots\ldots(79)\ bis,$$

and we shall obtain two values of r, r_1 and r_2, such that S_{r_1} contains W and S_{r_2} is contained by W. To simplify the procedure we shall make a linear transformation of coordinates independent of N, which will change S_{r_1} and S_{r_2} into hyperspheres with equations

$$N\sum_{s=1}^{k-1}(x_s^2)=r^2 \quad\ldots\ldots(85),$$

but will leave W as a hyperellipsoid.

The equations of transformation will be of the form

$$q_s-(\pi_s-\phi_s)=\beta_{s,1}x_1+\beta_{s,2}x_2+\ldots+\beta_{s,k-1}x_{k-1}\ (s=1, 2, \ldots k-1) \quad\ldots\ldots(86),$$

where of course certain conditions hold among the β's. The origin is now at the centre of the hyperspheres. If $\psi_1, \psi_2, \ldots \psi_{k-1}$ are the new coordinates of the centre of the transformed hyperellipsoid (79), then

$$\phi_s=\beta_{s,1}\psi_1+\beta_{s,2}\psi_2+\ldots+\beta_{s,k-1}\psi_{k-1} \quad\ldots\ldots(87),$$

so that every ψ will be a linear function of the ϕ's with constant coefficients, and therefore of the same order as the ϕ's.

Subtracting (87) from (86) we have,

$$q_s-\pi_s=\beta_{s,1}(x_1-\psi_1)+\ldots+\beta_{s,k-1}(x_{k-1}-\psi_{k-1}) \quad\ldots\ldots(88).$$

* That is to say as defined on p. 287.

If now we apply (88) to the transformation of the expression $N \overset{k}{\underset{s=1}{S}} \{(\pi_s - q_s)^2/p_s\}$ we see that since (86) applied to (84) gives (85), we shall obtain $N \overset{k-1}{\underset{s=1}{S}} (x_s - \psi_s)^2$. It follows that the transformed surface (79) may be written

$$N \overset{k}{\underset{s=1}{S}} \frac{(\pi_s - q_s)^2}{\pi_s} = N \overset{k}{\underset{s=1}{S}} \left\{ \frac{(\pi_s - q_s)^2}{p_s} \left(1 - \frac{\pi_s - p_s}{\pi_s}\right)\right\}$$

$$= N \overset{k-1}{\underset{s=1}{S}} (x_s - \psi_s)^2 + N\rho \underset{st}{S} \{a_{st}(x_s - \psi_s)(x_t - \psi_t)\} = \chi_1^2 \text{.....(89)},$$

where $\rho = \sqrt{\overset{k}{\underset{s=1}{S}} (\pi_s - p_s)^2}$, and the quantities a_{st} are functions of the π's, but do not depend on N and will tend to finite limits as $\pi_s \to p_s$ $(s = 1, 2, \ldots k)$.

We must now determine the radii r_1 and r_2 of the hyperspheres S_r which will contain and be contained by the transformed hyperellipsoid (89). Let P be some point $(x_1, x_2, \ldots x_{k-1})$ on (89), and let

$$d = \chi_1/\sqrt{N} + \eta$$

be its distance from the centre, Q, $(\psi_1, \psi_2, \ldots \psi_{k-1})$. Then $x_s - \psi_s = d \cos \gamma_s$, say, and

$$N\{d^2 + \rho d^2 \underset{st}{S} (a_{st} \cos \gamma_s \cos \gamma_t)\} = \chi_1^2,$$

or

$$d = \frac{\chi_1}{\sqrt{N}} + \eta = \frac{\chi_1}{\sqrt{N}} \frac{1}{\sqrt{1 + \rho \underset{st}{S} (a_{st} \cos \gamma_s \cos \gamma_t)}} = \frac{\chi_1}{\sqrt{N}} \{1 - \tfrac{1}{2}\rho \underset{st}{S} (a_{st} \cos \gamma_s \cos \gamma_t) + \ldots\}.$$

Hence $\eta = \chi_1 \rho \tau / \sqrt{N}$, where the value of τ depends upon the position of the point P on the hypersurface but is limited and independent of N.

We may now deduce certain inequalities with the help of the following diagram, in which O represents the origin, or centre of the hyperspheres (85); Q is the point $(\psi_1, \psi_2, \ldots \psi_{k-1})$ or centre of the hyperellipsoid (89); P is some point on this latter surface; $OP = R$; $OQ = g$; $PQ = d$.

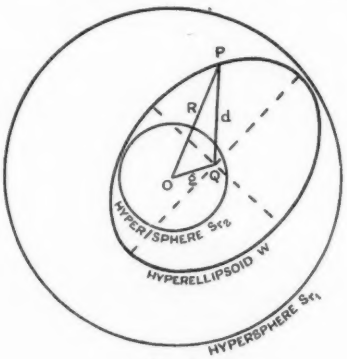


Fig. 6.

If S_{r_1} includes W, then $r_1/\sqrt{N}$ must be greater than R, a condition which is obviously sufficient. But $R < d + g$ and hence we may take for r_1 the upper limit of $(d+g)\sqrt{N}$, or

$$r_1 = \chi_1(1 + \rho\tau_1) + \sqrt{N \overset{k-1}{\underset{s=1}{S}} (\psi_s^2)} \text{ .....(90)},$$

where τ_1 is a finite number, the upper limit of $|\tau|$. In the same way if S_{r_2} is to be included in W, then it is necessary that $d > g + r_2/\sqrt{N}$, and so we may take for r_2 the lower limit of $(d-g)\sqrt{N}$, or

$$r_2 = \chi_1(1-\rho\tau_1) - \sqrt{N \sum_{s=1}^{k-1} (\psi_s^2)} \quad \text{.....(91).}$$

These conditions which hold for the transformed hypersurfaces will of course hold among the surfaces (79) and (84) before transformation. The values chosen for r_1 and r_2 depend upon the parameters α; we shall now find two constant quantities $\bar{r}_1$ and $\bar{r}_2$ which satisfy the above conditions throughout the region $\Omega(\chi_0)$.

Consider a region $\Omega(\lambda\chi_0)$ containing $\Omega(\chi_0)$, where λ is some constant independent of N which will be specified later. We shall now show that if the point $(\pi_1, \pi_2, \dots \pi_{k-1})$ lies within $\Omega(\lambda\chi_0)$, then the conditions regarding the envelopes will be satisfied within the smaller region $\Omega(\chi_0)$.

Since
$$\rho = \sqrt{\sum_{s=1}^{k} (p_s - \pi_s)^2},$$

and since in view of (81) within $\Omega(\lambda\chi_0)$

$$(p_s - q_s)^2 \leqslant p_s \chi_0^2 \lambda^2 / N,$$

it follows that

$$\rho \leqslant \frac{\lambda\chi_0}{\sqrt{N}} \sqrt{\sum_{s=1}^{k} (p_s)} = \frac{\lambda\chi_0}{\sqrt{N}} \quad \text{.....(92).}$$

But from the initial conditions regarding the locus of the π's (78), we know that if $\Omega(\lambda\chi_0)$ be small enough, that is to say if N be large enough, then for all points $(\pi_1, \pi_2, \dots \pi_{k-1})$ in this region the ϕ's, and therefore the ψ's, are of an order higher than ρ. We may therefore write

$$\sqrt{\sum_{s=1}^{k-1} (\psi_s^2)} = \zeta\rho < \lambda\chi_0\zeta_1/\sqrt{N} \quad \text{.....(93),}$$

where ζ_1 is a function depending on N but independent of the position of the point $(\pi_1, \pi_2, \dots \pi_{k-1})$ within $\Omega(\lambda\chi_0)$, which tends to zero as that region diminishes, or $N \to \infty$. Thus, provided that $(\pi_1, \pi_2, \dots \pi_{k-1})$ lies within $\Omega(\lambda\chi_0)$, it follows that for N sufficiently large,

$$\left.\begin{aligned} 0 < r_1 &\leqslant \chi_1 + \zeta_1\lambda\chi_0 + \lambda\chi_0\chi_1\tau_2/\sqrt{N} = \bar{r}_1 < \chi_0 \\ 0 < r_2 &\geqslant \chi_1 - \zeta_1\lambda\chi_0 - \lambda\chi_0\chi_1\tau_2/\sqrt{N} = \bar{r}_2 \end{aligned}\right\} \quad \text{.....(94),}$$

where τ_2 is a finite number, the upper limit of τ_1 for all systems of π's corresponding to points inside $\Omega(\chi_0)$. We have now found an upper limit to r_1 and a lower limit to r_2 which hold good for all points $(\pi_1, \pi_2, \dots \pi_{k-1})$ within $\Omega(\lambda\chi_0)$, and it follows that if we give to r in (84) these two constant values, the hypersurface $S_{\bar{r}_1}$ will always contain and $S_{\bar{r}_2}$ always be contained by the corresponding hypersurface W of (79). It is easy to see that if this be so then the envelope $\mathcal{E}'$ of $S_{\bar{r}_1}$ will lie completely outside the envelope E of W, and that the latter will lie completely outside the envelope $\mathcal{E}''$ of $S_{\bar{r}_2}$, for that portion of E corresponding to points $(\pi_1, \pi_2, \dots \pi_{k-1})$ lying within $\Omega(\lambda\chi_0)$. This may be proved as follows.

Suppose $\mathcal{E}'$ to cut E, and that P is one of the points of intersection in the transformed space. P must belong to some hypersphere $S_{\bar{r}_1}'$ with centre at C' and to some hyperellipsoid W'. We have established a correspondence between hyperellipsoids W and hyperspheres $S_{\bar{r}_1}$, but obviously W' cannot correspond to $S_{\bar{r}_1}'$, or they would not cut at P. Let $S_{\bar{r}_1}''$ be the hypersphere corresponding to W'; it will contain it entirely and thus also the point P. Hence if C'' be the centre of $S_{\bar{r}_1}''$, it follows that $C''P < \bar{r}_1$. But if L be the locus of centres of $S_{\bar{r}_1}$, then the nearest point on L to P must be C', the centre of $S_{\bar{r}_1}'$, and this shortest distance must equal $\bar{r}_1$. But C'' also lies on L and we have seen that $C''P < \bar{r}_1$. We have here a contradiction; in fact, $\mathcal{E}'$ cannot cut E. An analogous proof establishes the fact that $\mathcal{E}''$ cannot cut E; in this case we consider the values of χ^2 corresponding to a supposed point of intersection P and different centres of the hyperellipsoids W.

It remains to show that λ may be so chosen that $\Omega(\chi_0)$ will contain no points on hypersurfaces W whose centres $(\pi_1, \pi_2, \dots \pi_{k-1})$ lie outside $\Omega(\lambda\chi_0)$. For if this can be done, then the required conditions regarding $\mathcal{E}'$, E and $\mathcal{E}''$ will hold throughout the region $\Omega(\chi_0)$. In view of (79), for every point on W

$$|\pi_s - q_s| \leqslant \chi_1 \sqrt{\pi_s/N} < \chi_1/\sqrt{N} \quad \dots\dots(95),$$

and therefore the distance of such a point from the centre $(\pi_1, \pi_2, \dots \pi_{k-1})$ of W is always $\leqslant \chi_1\sqrt{(k-1)/N}$. The limiting surfaces of $\Omega(\lambda\chi_0)$ and $\Omega(\chi_0)$ are similar and coaxial hyperellipsoids, the former being an enlargement of the latter in the ratio of λ to 1. It follows that the minimum distance between the two surfaces is $\mu(\lambda-1)\chi_0/\sqrt{N}$, where μ is some quantity depending on the shape of the hyperellipsoids but independent of N. But λ may be chosen so that

$$\mu(\lambda-1)\chi_0/\sqrt{N} > \sqrt{k-1}\,\chi_1/\sqrt{N} \quad \dots\dots(96)$$

independently of N. If this choice has been made, then $\Omega(\chi_0)$ will include only the whole or parts of hypersurfaces W whose centres lie within $\Omega(\lambda\chi_0)$, and as we have seen for all such surfaces the envelope E of W lies between the envelopes $\mathcal{E}'$ and $\mathcal{E}''$ of the surfaces $S_{\bar{r}_1}$ and $S_{\bar{r}_2}$.

Denote by AB the common part of any two regions A and B. Then

$$C\iint\dots\int_{\mathcal{E}''\Omega(\chi_0)} e^{-\frac{1}{2}\chi^2} dq_1 \dots dq_{k-1} < C\iint\dots\int_{E\Omega(\chi_0)} e^{-\frac{1}{2}\chi^2} dq_1 \dots dq_{k-1}$$
$$< C\iint\dots\int_{\mathcal{E}'\Omega(\chi_0)} e^{-\frac{1}{2}\chi^2} dq_1 \dots dq_{k-1} \quad \dots\dots(97),$$

and further since from condition (83) the integral of $Ce^{-\frac{1}{2}\chi^2}$ taken outside $\Omega(\chi_0)$ is less than ϵ, it follows that

$$C\iint\dots\int_{\mathcal{E}''} e^{-\frac{1}{2}\chi^2} dq_1 \dots dq_{k-1} - \epsilon < C\iint\dots\int_{E} e^{-\frac{1}{2}\chi^2} dq_1 \dots dq_{k-1}$$
$$< C\iint\dots\int_{\mathcal{E}'} e^{-\frac{1}{2}\chi^2} dq_1 \dots dq_{k-1} + \epsilon \quad \dots\dots(98).$$

But the values of the first and third integrals in (98) are known, being taken inside the envelopes of the hypersurfaces

$$N \sum_{s=1}^{k} \frac{(\pi_s - \phi_s - q_s)^2}{p_s} = \bar{r}_1^2, \qquad N \sum_{s=1}^{k} \frac{(\pi_s - \phi_s - q_s)^2}{p_s} = \bar{r}_2^2 \quad \dots\dots(99),$$

where p_s and $\bar{r}_1$ and $\bar{r}_2$ are constant, these equations (99) corresponding in view of (78) to equations (62) of our Lemma. We have in fact from (63) that if N is sufficiently large

$$C\int_0^{\bar{r}_2} \chi^{k-c-2} e^{-\frac{1}{2}\chi^2} d\chi - \epsilon < C\iint\dots\int_E e^{-\frac{1}{2}\chi^2} dq_1 \dots dq_{k-1} < C\int_0^{\bar{r}_1} \chi^{k-c-2} e^{-\frac{1}{2}\chi^2} d\chi + \epsilon \quad \dots\dots(100),$$

where
$$1/C = \int_0^{\infty} \chi^{k-c-2} e^{-\frac{1}{2}\chi^2} d\chi.$$

Thus

$$-C\int_{\bar{r}_2}^{\chi_1} \chi^{k-c-2} e^{-\frac{1}{2}\chi^2} d\chi - \epsilon < C\iint\dots\int_E e^{-\frac{1}{2}\chi^2} dq_1 \dots dq_{k-1} - C\int_0^{\chi_1} \chi^{k-c-2} e^{-\frac{1}{2}\chi^2} d\chi$$
$$< C\int_{\chi_1}^{\bar{r}_1} \chi^{k-c-2} e^{-\frac{1}{2}\chi^2} d\chi + \epsilon \quad \dots\dots(101).$$

But we know from (94) that as $N \to \infty$

$$\mathrm{Lt}\,(\bar{r}_1) = \mathrm{Lt}\,(\bar{r}_2) = \chi_1,$$

and hence we see that ϵ being arbitrarily small,

$$\mathop{\mathrm{Lt}}_{N\to\infty} C\iint\dots\int_E e^{-\frac{1}{2}\chi^2} dq_1 \dots dq_{k-1} = C\int_0^{\chi_1} \chi^{k-c-2} e^{-\frac{1}{2}\chi^2} d\chi \quad \dots\dots(102),$$

which establishes the Theorem set out on p. 291 above.

THE CONTRIBUTION OF GIOVANNI PLANA TO THE NORMAL BIVARIATE FREQUENCY SURFACE.

By KARL PEARSON.

Miss Helen M. Walker has contributed a very interesting note on "The Relation of Plana and Bravais to the Theory of Correlation" in the June issue of *Isis**. I am concerned here only with Plana's memoir of 1812—"Mémoire sur divers problèmes de probabilité"†. Miss Walker has, I venture to think, at the same time under-estimated and over-estimated Plana's contributions to the subject. The under-estimation arises from her misreading of what Plana is doing; a misreading to some extent justified by his vague language and awkward choice of symbols. Her over-estimation arises from slips in her own algebra, when she comes to interpret in modern symbols Plana's results.

Let us consider generally what Plana is dealing with. Suppose $\epsilon_1, \epsilon_2, \ldots \epsilon_p$ to be p independent variables, and $q_1, q_2, \ldots q_p$ one system of purely numerical factors, and $q_1', q_2', \ldots q_p'$ a second like system, then Plana introduces two dependent variates Q and Q' as linear functions of the independent variates $\epsilon_1, \epsilon_2, \ldots \epsilon_p$:

$$Q = q_1 \epsilon_1 + q_2 \epsilon_2 + \ldots + q_p \epsilon_p,$$
$$Q' = q_1' \epsilon_1 + q_2' \epsilon_2 + \ldots + q_p' \epsilon_p.$$

Now if the ϵ's be measured from their means, and their variances, σ_ϵ^2, be all the same, and σ denote as usual a standard deviation, we have since mean Q = mean $Q' = 0$:

$$\sigma_Q^2 = (q_1^2 + q_2^2 + \ldots + q_p^2)\ \sigma_\epsilon^2 = A \,.\, \sigma_\epsilon^2,$$
$$\sigma_{Q'}^2 = (q_1'^2 + q_2'^2 + \ldots + q_p'^2)\ \sigma_\epsilon^2 = C \,.\, \sigma_\epsilon^2,$$
$$\sigma_Q \sigma_{Q'} r_{QQ'} = (q_1 q_1' + q_2 q_2' + \ldots + q_p q_p')\ \sigma_\epsilon^2 = B \,.\, \sigma_\epsilon^2,$$

and accordingly:

$$r_{QQ'} = B/\sqrt{AC}.$$

The symbols A, B, C are those of Plana, but the method of finding the relation of Q and Q', and the correlation coefficient have nothing to do with him. Now let us assume for the moment that the surface is normal then we have for its equation:

$$z = \frac{N}{2\pi\sigma_Q\sigma_{Q'}\sqrt{1-r_{QQ'}^2}}\, e^{-\frac{1}{2}\frac{1}{1-r_{QQ'}^2}\left(\frac{Q^2}{\sigma_Q^2} - \frac{2QQ'r_{QQ'}}{\sigma_Q\sigma_{Q'}} + \frac{Q'^2}{\sigma_{Q'}^2}\right)}.$$

* Vol. x. pp. 466—484.

† *Mémoires de l'Académie impériale des Sciences...de Turin*, 1811—1812, Vol. xx. pp. 355—408, Torino, 1813.

Substituting for $r_{QQ'}$ and σ_Q, $\sigma_{Q'}$ we have

$$z = \frac{N}{2\pi\sigma_\epsilon^2\sqrt{E}} e^{-\frac{1}{2}\frac{1}{E\sigma_\epsilon^2}(CQ^2 - 2BQQ' + AQ'^2)},$$

where we write with Plana $E = AC - B^2$ (p. 380), or taking $\sigma_\epsilon^2 = 2a$:

$$z = \frac{N}{4\pi a\sqrt{E}} e^{-\frac{1}{2}\frac{1}{2Ea}(CQ^2 - 2BQQ' + AQ'^2)},$$

which is identical with Plana's Equation (α') on p. 380.

Now Plana does not deduce this in its above general form. On the other hand, he does not assume the surface to be normal. He supposes that ϵ is obtained from throwing a $2n$ faced die, whose faces carry the numbers:

$$n,\ n-1,\ n-2,\ \ldots,\ 1,\ -1,\ -2,\ \ldots,\ -n,$$

and his p ϵ's are all obtained from uniform similar $2n$ faced dice. Accordingly it follows that his $\sigma_\epsilon^2 = \frac{1}{6}(n+1)(2n+1)$, or his $a = \frac{1}{12}(n+1)(2n+1)$. Substituting this we have:

$$z = \frac{3N}{\pi(n+1)(2n+1)\sqrt{E}} e^{-\frac{3(CQ^2 - BQQ' + AQ'^2)}{E(n+1)(2n+1)}},$$

which is Plana's result at the top of p. 381 *.

How does Plana reach his normal surface? He does it in the manner by which Laplace deduces the normal curve. He is assuming p contributory cause-groups each of which may give $2n$ equally likely half-positive and half-negative deviations. If s be an integer, he has no hesitation in writing $\frac{s}{n} = x$ and $\frac{1}{n} = dx$ in order to pass from a summation to an integral (p. 376)†. Also in the manner of Laplace, he expands one variable in powers of a second, possibly only to two terms, and if the second term be of the order $\frac{1}{p}$, where p is the number of dice, he does not hesitate to neglect not only that term but the sum of all the remaining terms‡. Some of his processes of integration would hardly now-a-days be considered legitimate without further investigation. Still he deduces the bivariate normal surface, on hypotheses of precisely the same nature as Laplace used to deduce the normal curve.

There is, however, something which, I think, Miss Walker has overlooked. Plana deduces his correlation surface for Q and Q', by supposing them to be linear functions of p independent variables; or he supposes the relationship of Q and Q' to arise from their containing at least some common cause groups. It is true that he supposes all these cause groups to be of the same character, i.e. similar dice; further he considers his cause group to be such that all the values that

* Plana puts the total number $(2n)^p$ of combinations of faces of the p dice in place of our N.

† This seems to suggest that n as well as p is to be considered large.

‡ Compare for example his procedure on p. 375. If justifiable it is certainly not there justified.

occur are equally likely to occur. He is approximating to a continuous rectangular frequency for his individual ϵ. We might now-a-days assume it to be better given by a normal curve with a different standard deviation for each ϵ, and the proof of the bivariate frequency distribution then becomes rigid not roughly approximate. It is well known that if we have p frequency distributions, $\phi_1(x_1)$, $\phi_2(x_2)$, ... $\phi_p(x_p)$ of any kind and sample at random one member of each, and find the weighted mean of these sampled members, the distribution of such means approaches closer and closer to the normal curve as we increase the number p of our $\phi_s(x_s)$ distributions. It should not be hard to show that if we take two different systems of weighting for each system of sampled members, the two means would approximately follow the distribution determined by the bivariate normal surface. This would be the satisfactory generalisation of Plana's Theorem.

Now let us turn back to Miss Walker's paper. On p. 474, she writes down Plana's expressions for Q and Q' correctly supplying his missing dash to the second Q and also correcting his second $q^{(p)}$ and $\epsilon_{(p)}$—then she says:

> The part of the paper in which we are primarily interested at present begins on page 376 with a question as to what is the probability that the two equations:
>
> $$q'\epsilon'+q''\epsilon''+q'''\epsilon'''+\ldots+q^{(p)}\epsilon^{(p)}=Q,$$
> $$q_{,}\epsilon'+q_{,,}\epsilon''+q_{,,,}\epsilon'''+\ldots+q_{(p)}\epsilon^{(p)}=Q'^{*},$$
>
> shall be satisfied simultaneously. This is only another way of asking what is the probability that in one throw of $P[?\,p]$ dice the sum Q shall be exposed, while in another throw of the same number of similar dice, the sum exposed shall be Q'. (*Isis*, Vol. x. p. 474.)

Now this statement that it is "another way of asking, etc." is wholly erroneous and does not apply to Plana's analysis. There is only *one* throw of the dice and the two Q's are obtained by different weightings; $q^{(p)}$ is not $q_{(p)}$. With *two* throws of the dice, as Miss Walker supposes, there would be no correlation between Q and Q'. Here she misreads what Plana is doing and underestimates what he has achieved, namely the idea of making two variates inter-related by supposing them linear functions of the *same* variates. Such inter-relations between directly observed and dependent variates had arisen in geodesic work, for example when three sides of a triangle were measured and the angles determined from these values of the sides by the trigonometrical equations, then the error of the angle A was correlated with that of the angle B, and such cases were probably in Plana's mind. But here he is seeking a relation between Q and Q', based it is true on a somewhat narrow assumption as to the common cause groups on which they depend, but nevertheless with some appreciation that the dependence is the product of common factors.

If we turn from the case of under-estimation to that of over-estimation we have to consider Miss Walker's discussion on p. 475. She compares Plana's

$$z=\frac{N}{4\pi a\sqrt{E}}e^{-\frac{1}{4aE}(CQ^2-2BQQ'+AQ'^2)},$$

* Plana has incorrectly $+q^{(p)}\epsilon_{(p)}=Q$ for the last terms of this line.

with the familiar form

$$z = \frac{N}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}}\, e^{-\frac{1}{2(1-r^2)}\left(\frac{x^2}{\sigma_1^2} - \frac{2rx_1x_2}{\sigma_1\sigma_2} + \frac{x_2^2}{\sigma_2^2}\right)},$$

and says:

"We see that $E = 1 - r^2$, $C = \frac{\sigma_2}{\sigma_1}$, $A = \frac{\sigma_1}{\sigma_2}$, $a = \frac{1}{2}\sigma_1\sigma_2$, and $B = r$."

Unfortunately every one of these equations is in error! Miss Walker has failed to observe that in the first equation there are five constants a, E, C, B and A, while in the second there are only three σ_1, σ_2 and r. Accordingly if it were legitimate to assume values of the five which lead to the second equation, there is an endless variety that could be chosen. Actually the values of A, B, C, E and a have already been fixed in the course of Plana's work and cannot be varied at choice. E has been written by Plana (p. 380) to briefly express $AC - B^2$, the values of A, B, C being those given by me above. Further the value of a is $\frac{1}{12}(n+1)(2n+1)$ and the value of $r = \frac{B}{\sqrt{AC}}$. Thus we have:

$$E = AC(1-r^2),\ C = \frac{\sigma_2}{\sigma_1}\sqrt{AC},\ A = \frac{\sigma_1}{\sigma_2}\sqrt{AC},\ a = \tfrac{1}{2}\sigma_1\sigma_2\sqrt{AC} \text{ and } r = B/\sqrt{AC}.$$

These equations are incompatible with Miss Walker's unless $AC = 1$. But this involves the special relation:

$$(q_1^2 + q_2^2 + \ldots + q_p^2)(q_1'^2 + q_2'^2 + \ldots + q_p'^2) = 1.$$

But no such relation actually exists. Plana indeed writes on his p. 372:

"L'on suppose q', q'', q''', $\ldots q^{(p)}$ nombres entiers," which puts any such permanent relation out of the question. Personally I do not see any real necessity for this restriction, but there is a great improbability of such a relation as $AC = 1$ occurring even if the q's were fractional. Miss Walker making $AC = 1$ concludes (p. 476) that Plana's "was the only use of a single symbol equivalent to our coefficient of correlation in the equation for the surface of normal probability" before the time of Galton. And again "Plana's formula has in it a single letter B, which exactly takes the place of r in the formula for the surface of normal correlation." (Summary, p. 483.)

This is an over-statement of the case due to faulty algebra. $B/\sqrt{AC}$, and not B, takes the place of r, and Plana does not refer to the former expression, still less understand its significance for association. But he has reached as a particular case, that of throwing an indefinitely large number of dice, and considering two differently weighted means of their results, a solution which gives *roughly* a normal surface with a product term. If we are merely looking for a normal surface obtained from any hypothesis and having product terms, then Lagrange reached such a surface in 1770, for $m - 1$ variates, and no doubt Plana was well acquainted with his wife's uncle's writings.

THE RELATION OF WEIGHT TO STATURE, BI-CRISTAL DIAMETER AND AGE*.

By HORACE GRAY, M.D.

Introductory.

What is the normal weight for a given individual? The problem is to improve present methods of prediction, because all of them give such wide variabilities. Indices like weight divided by stature, or by stature squared or cubed, will not be considered here; but rather tables of double-entry, such as the common standards of weight for known stature and year of age.

In view of the increasingly common practice among pediatricians of predicting weight (nutrition) not primarily from age but from stature (bone), or better from height together with age, an obvious line of study is the replacement of the age factor by some measurement of body width.

This premise has been pursued in previous studies, and has also been studied by Root and Miles (1, 2, 3). Chest circumference was tried first, next the transverse and antero-posterior diameters of the chest, but all were found wanting (4).

Recently (5), by use of more refined statistical methods, an analysis was made of the degree of relationship between weight (W), age (A), stature (S), and particularly various nearly bony diameters: bi-acromial breadth of the shoulders (Ba), bi-cristal breadth of pelvic brim (Bc), maximal bi-styloid diameter of wrist (Bs), and maximal bi-malleolar diameter of ankle (Bm). The material consisted of private-school boys aged 4 to 20 years, to the number of 810, except that for the diameters of wrist and ankle the populations were 290 to 293 respectively. The resulting correlation coefficients may be worth a brief review, since the zero order coefficients have been published only in society proceedings, while the more recently calculated zero order correlation between weight and bi-acromial diameter and all the partial coefficients of correlation have not been printed at all.

If other things be equal, that trait most highly correlated with weight might be expected to be the most dependable for predicting weight. Accordingly if we note in Table I the relationship of weight to various traits, this was highest for stature, then bi-cristal, then bi-acromial, then bi-styloid, then age, and finally bi-malleolar diameter. Furthermore the difference between r_{WA} and r_{WS} proved to be statistically significant, the difference being $\cdot 0379 \pm \cdot 0036$, with a ratio of the difference to its probable error of 10·4. Likewise the difference between r_{WA} and r_{WBc}, namely, $\cdot 0240 \pm \cdot 0045$, with a ratio of the difference to the probable error of the difference

* This is one of a series of articles by members of the staff of the Behaviour Research Fund, Illinois Institute for Juvenile Research, Chicago, Herman M. Adler, M.D., Director. Series B, Number 135.

TABLE I.

Correlation Coefficients, Zero Order, in Order of Magnitude.

Stature and Age	·9320 ± ·0031
Weight and Stature	·9268 ± ·0033
Bi-cristal and Stature ...	·9234 ± ·0035
Weight and Bi-cristal ...	·9129 ± ·0039
Weight and Bi-acromial ...	·9082 ± ·0042
Bi-styloid and Stature ...	·9028 ± ·0073
Bi-malleolar and Stature ...	·8971 ± ·0077
Weight and Bi-styloid ...	·8936 ± ·0080
Weight and Age	·8889 ± ·0050
Bi-cristal and Age	·8700 ± ·0058
Weight and Bi-malleolar ...	·8606 ± ·0102
Bi-styloid and Age	·8576 ± ·0105
Bi-malleolar and Age ...	·8334 ± ·0120

of 5·3, was significant. But the correlation of weight with bi-styloid diameter of the wrist, although larger than r_{WA}, was not significantly larger, the difference being ·0047 ± ·0070. On comparing the relative value of bi-cristal and bi-acromial diameters, it was found that the difference between r_{WBc} and r_{WBa} was ·0047 ± ·0044, that is, not statistically significant. Moreover, the bi-acromial breadth is difficult to measure accurately in that duplicate measurements on the same subject by the same measurer at the same time not infrequently disagree by 10 mm., owing to the subject's squaring his shoulders by shifts which are often barely perceptible to the measurer's eye, while the bi-cristal breadth is one of the most precise measurements available on the human body in that repeated measurements agree within 1 mm.

TABLE II.

Partial Correlation Coefficients

First Order		Second Order	
$r_{S,A.Ba}$ ...	·7134 ± ·0196	$r_{W,S.ABa}$ ...	·5987 ± ·0253
$r_{Bc,S.A}$ ...	·6301 ± ·0143	$r_{W,Bc.AS}$ ...	·3916 ± ·0200
$r_{W,Bc.A}$ ...	·6181 ± ·0146	$r_{W,Ba.AS}$ ...	·3410 ± ·0350
$r_{W,S.A}$ ...	·5920 ± ·0153	$r_{W,S.ABc}$ ...	·3317 ± ·0211
$r_{W,Ba.A}$ ...	·5573 ± ·0272	$r_{W,A.SBc}$ ...	·2724 ± ·0220
$r_{Ba,S.A}$ ...	·5552 ± ·0272	$r_{W,A.SBa}$ ...	·1918 ± ·0382
$r_{W,S.Bc}$ ...	·5349 ± ·0168		
$r_{W,A.Ba}$ ...	·5306 ± ·0285		
$r_{W,Ba.A}$ ...	·4732 ± ·0305		
$r_{W,A.Bc}$ ...	·4706 ± ·0185		
$r_{W,Bc.S}$ ...	·3947 ± ·0200		
$r_{W,Ba.S}$ ...	·3523 ± ·0347		
$r_{W,A.S}$ ...	·1844 ± ·0230		
$r_{W,Ba.S}$ ...	·1760 ± ·0383		

for the same observer, and even within 2 mm. for different observers. In sum it was concluded that weight is most nearly correlated with stature, next best to bi-cristal diameter, but only fifth best to age.

Other things, however, are not equal, and therefore partial correlation coefficients are desirable in order to render constant those variables which are likely to affect the underlying correlation. In Table II, the first order partials showed that, when age was constant, weight was slightly more highly correlated with bi-cristal diameter ($r = \cdot6181 \pm \cdot0146$) than with stature ($r = \cdot5920 \pm \cdot0153$), whereas when stature was fixed weight was more highly correlated with bi-cristal diameter ($\cdot3947 \pm \cdot0200$) than with age ($\cdot1844 \pm \cdot0230$). The second order partials showed that weight was distinctly more highly correlated with bi-cristal diameter when stature and age were fixed ($r = \cdot3916 \pm \cdot0200$) than with age when stature and bi-cristal were constant ($r = \cdot2724 \pm \cdot0220$). Hence it was expected that the multiple correlation of weight with stature and pelvic diameter would be better than that with stature and age, and in fact the multiple correlation coefficient R of weight on various pairs of traits, each pair containing stature, was

$$R_{W.SBc} = \cdot9435 \pm \cdot0026,$$
$$R_{W.SBs} = \cdot9362 \pm \cdot0049,$$
$$R_{W.SBm} = \cdot9292 \pm \cdot0054,$$
$$R_{W.SA} = \cdot9294 \pm \cdot0032.$$

Furthermore, the difference between the first and last of these multiple correlation coefficients was statistically significant, being ($\cdot0141 \pm \cdot0015$), nine times its probable error. The difference between the second and last coefficients was $\cdot0068 \pm \cdot0027$, hardly significant. In the paper reviewed, therefore, the main inference seemed that weight is less closely related to age than to certain body diameters.

PRESENT STUDY.

A grave objection which may be raised to the foregoing analysis is the apparent non-linearity of some of the distributions. Is this marked enough to upset the main inference?

Subjects Studied. For the present paper the extreme ages with their small number of subjects were omitted, namely, central ages 4, 18, 19 and 20. Otherwise the material was that used in the paper just summarised, to wit, boys in the following six private schools: the Alcott School in Lake Forest, Illinois, the North Shore Country Day School in Winnetka, Illinois, the Francis W. Parker School in Chicago, the Cate School and the Deane School in Santa Barbara, California, and Groton School in Massachusetts. The racial strain may be defined as North-west European, mainly of English-speaking stocks, including a few Jewish boys. From the viewpoint of normal build, those boys were omitted who were markedly obese, subject to shortness of breath, coughs or colds, or otherwise questionable as to health; the omissions totalled about five per cent. of the gross number measured, leaving 733 boys for this study.

Anthropological Technique. Weight and all measurements were taken without clothes, usually between 9 a.m. and 3 p.m. Stature was measured against a wall, the larger diameters with Hrdlička's large sliding caliper with straight arms of metal, and the wrist and ankle diameters with his small metal sliding caliper (6).

Statistical Results. The distributions are shown in Tables IV—IX.

The observed regression lines y on x, obtained by splining through the plotted means of weights (y) for given columns (x), such as bi-cristal diameter or age, are shown in Figs. i to vi. These would seem to be the regressions useful for predicting weight; the reverse predictions, i.e. from weight, do not interest us here, and therefore have not been plotted.

The calculated constants are shown in Tables III A and III B.

TABLE III A.

Means and Standard Deviations of Characters measured on 733 *Boys.*

	Mean ± P.E.	S.D. ± P.E.
Age in years	12·416 ± 0·081	3·259 ± 0·057
Weight in kilos ...	43·810 ± 0·360	14·464 ± 0·255
Stature in cms. ...	153·140 ± 0·456	18·680 ± 0·329
Bi-cristal Diam. in mms.	241·301 ± 0·732	29·372 ± 0·517

TABLE III B.

Association between the Variables, in Order of Magnitude of the Zero Order Correlation Coefficients.

Variables		Correlation Coefficient	Correlation Ratio	Linearity	Criterion	Correlation Ratio	Linearity	Criterion
y	x	r	$\eta_{y.x}$	$\eta^2 - r^2 = \zeta$	ζ/P.E.	$\eta_{x.y}$	$\eta^2 - r^2$	—
Stature	Age	·9338 ± ·0032	·9351 ± ·0031	·0025 ± ·0025	1·00	·9414 ± ·0028	·0143 ± ·0060	2·40
Weight	Stature	·9313 ± ·0033	·9445 ± ·0027	·0247 ± ·0078	3·16	·9556 ± ·0022	·0457 ± ·0106	4·32
Bi-cristal	Stature	·9262 ± ·0035	·9284 ± ·0034	·0041 ± ·0024	1·71	·9351 ± ·0031	·0166 ± ·0047	3·53
Weight	Bi-cristal	·9136 ± ·0041	·9187 ± ·0039	·0093 ± ·0048	1·94	·9270 ± ·0035	·0246 ± ·0078	3·16
Bi-cristal	Age	·8753 ± ·0058	·8780 ± ·0057	·0048 ± ·0034	1·39	·8958 ± ·0049	·0365 ± ·0094	3·86
Weight	Age	·8732 ± ·0059	·8815 ± ·0056	·0146 ± ·0060	2·44	·9105 ± ·0043	·0665 ± ·0127	5·25

Now, recollecting the principle that "Variables highly correlated with the predicate and lowly correlated with each other is the ideal system to be hunted for" (7), we find in fact that the bi-cristal diameter is more highly correlated with the predicate weight than is age, the difference between r_{WBc} and r_{WA} being significant: ·0404 ± ·0051; likewise bi-cristal diameter is somewhat more lowly correlated with the other variable stature than is age, though here the difference, r_{AS} minus r_{SBc}, is not statistically significant: ·0076 ± ·0042.

The multiple correlation coefficient R of weight on the rival pairs of traits, each pair containing stature, is as follows:

$$R_{W.SBc} = ·9411 \pm ·0029,$$
$$R_{W.SA} = ·9314 \pm ·0033.$$
$$\text{Difference} = ·0097 \pm ·0018.$$

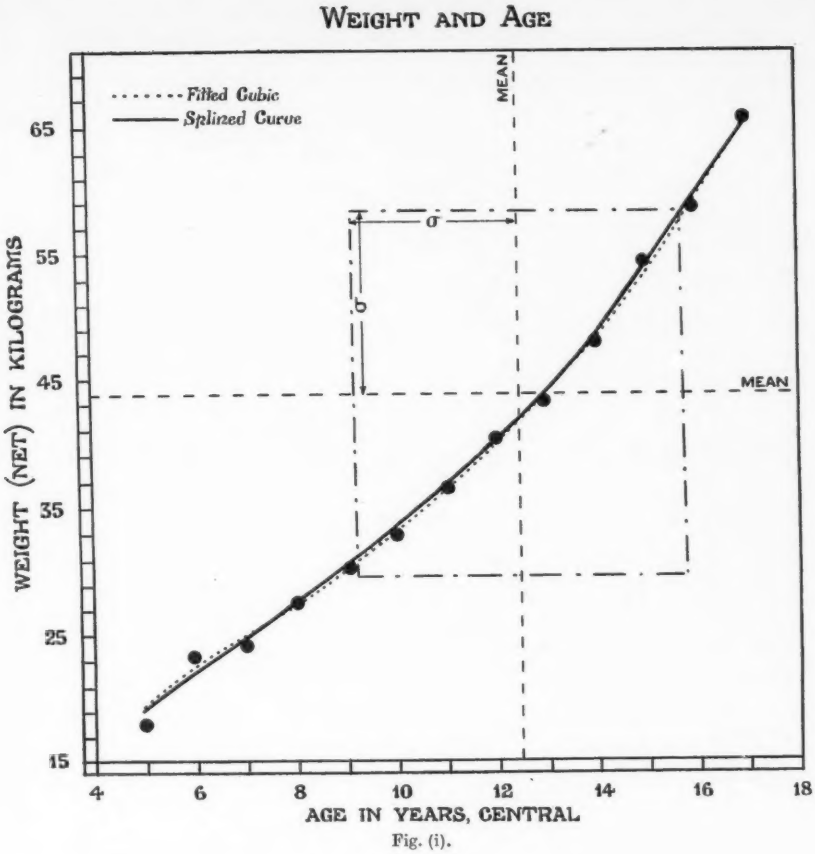


Fig. (i).

TABLE IV.

Weight and Age.

Age in Years, Central.

Weight (net) in Kilograms.	5	6	7	8	9	10	11	12	13	14	15	16	17	Totals
95·0—98·9	—	—	—	—	—	—	—	—	—	—	—	—	1	1
91·0—94·9	—	—	—	—	—	—	—	—	—	—	—	—	—	0
87·0—90·9	—	—	—	—	—	—	—	—	—	—	—	—	—	0
83·0—86·9	—	—	—	—	—	—	—	—	—	—	—	—	2	2
79·0—82·9	—	—	—	—	—	—	—	—	—	—	—	1	4	5
75·0—78·9	—	—	—	—	—	—	—	—	—	—	2	—	4	6
71·0—74·9	—	—	—	—	—	—	—	—	—	—	3	—	4	7
67·0—70·9	—	—	—	—	—	—	—	—	—	—	3	7	9	19
63·0—66·9	—	—	—	—	—	—	—	—	—	1	7	16	13	37
59·0—62·9	—	—	—	—	—	—	—	—	1	7	15	19	13	55
55·0—58·9	—	—	—	—	—	—	1	1	2	4	17	12	6	43
51·0—54·9	—	—	—	—	—	—	—	1	5	13	18	11	8	56
47·0—50·9	—	—	—	—	—	—	2	8	10	20	13	7	1	61
43·0—46·9	—	—	—	—	—	1	5	12	20	21	8	4	—	71
39·0—42·9	—	—	—	—	—	3	10	21	23	8	4	1	—	70
35·0—38·9	—	—	—	1	6	15	10	16	9	4	6	1	—	68
31·0—34·9	—	—	2	4	17	16	21	7	7	3	1	—	—	78
27·0—30·9	—	5	5	12	15	11	4	6	—	—	—	—	—	58
23·0—26·9	1	7	19	13	10	4	1	—	—	—	—	—	—	55
19·0—22·9	5	11	11	2	—	—	1	—	—	—	—	—	—	30
15·0—18·9	8	1	1	—	—	—	—	—	—	—	—	—	—	10
11·0—14·9	1	—	—	—	—	—	—	—	—	—	—	—	—	1
Totals	15	24	38	32	48	50	55	72	77	81	97	79	65	733

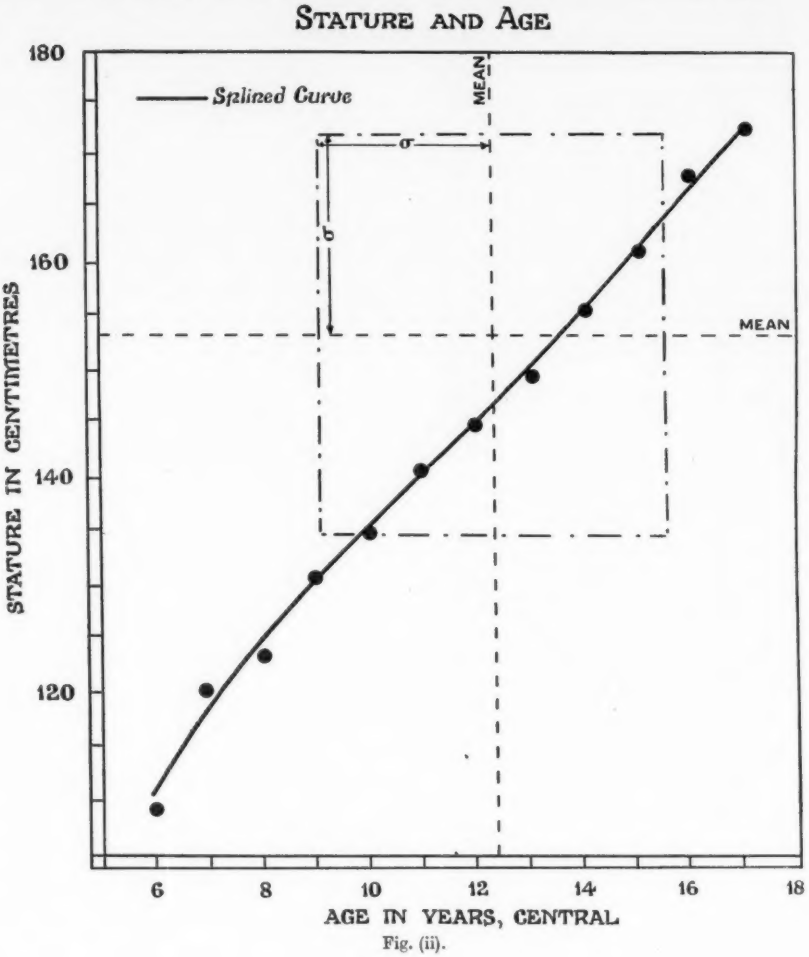


Fig. (ii).

TABLE V.

Stature and Age.

Age in Years, Central.

Stature in Centimetres.	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	Totals
190·5—195·4	—	—	—	—	—	—	—	—	—	—	—	1	1	2
185·5—190·4	—	—	—	—	—	—	—	—	—	—	2	—	5	7
180·5—185·4	—	—	—	—	—	—	—	—	—	—	3	11	10	24
175·5—180·4	—	—	—	—	—	—	—	—	—	—	11	16	20	47
170·5—175·4	—	—	—	—	—	—	—	—	—	7	23	24	21	75
165·5—170·4	—	—	—	—	—	—	1	1	4	16	22	15	6	65
160·5—165·4	—	—	—	—	—	—	—	—	14	20	20	8	1	63
155·5—160·4	—	—	—	—	—	—	2	10	21	26	9	4	1	73
150·5—155·4	—	—	—	—	—	4	8	28	22	7	6	—	—	75
145·5—150·4	—	—	—	—	1	9	17	18	14	5	1	—	—	65
140·5—145·4	—	—	—	1	5	12	16	7	2	—	—	—	—	43
135·5—140·4	—	—	1	4	18	14	8	7	—	—	—	—	—	52
130·5—135·4	—	1	4	12	15	9	2	1	—	—	—	—	—	44
125·5—130·4	—	5	9	11	8	1	—	—	—	—	—	—	—	34
120·5—125·4	—	5	16	4	1	1	—	—	—	—	—	—	—	27
115·5—120·4	1	8	5	—	—	—	—	—	—	—	—	—	—	14
110·5—115·4	5	4	1	—	—	—	1	—	—	—	—	—	—	11
105·5—110·4	8	1	2	—	—	—	—	—	—	—	—	—	—	11
100·5—105·4	1	—	—	—	—	—	—	—	—	—	—	—	—	1
Totals	15	24	38	32	48	50	55	72	77	81	97	79	65	733

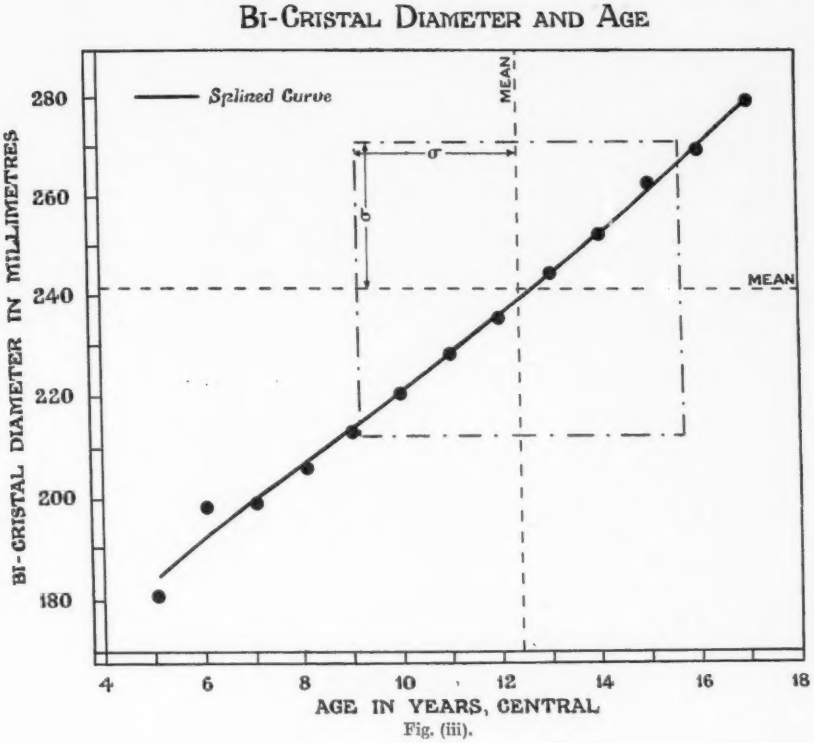


Fig. (iii).

TABLE VI.

Bi-cristal Diameter and Age.

Age in Years, Central.

Bi-cristal Diameter in Millimetres.	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	Totals
315—324	—	—	—	—	—	—	—	—	—	—	—	2	1	3
305—314	—	—	—	—	—	—	—	—	—	—	—	—	3	3
295—304	—	—	—	—	—	—	—	—	—	—	2	2	5	9
285—294	—	—	—	—	—	—	—	—	—	1	6	9	15	31
275—284	—	—	—	—	—	—	1	1	2	4	13	18	19	58
265—274	—	—	—	—	—	—	—	—	5	14	23	18	12	72
255—264	—	—	—	—	—	—	1	1	11	20	24	15	7	79
245—254	—	—	—	—	—	—	5	17	22	20	17	9	3	93
235—244	—	—	—	—	—	7	14	22	18	15	9	4	—	89
225—234	—	—	—	—	5	13	7	18	13	4	3	2	—	65
215—224	—	3	2	5	18	16	20	9	6	3	—	—	—	82
205—214	—	5	11	11	17	10	5	4	—	—	—	—	—	63
195—204	2	7	13	15	8	3	1	—	—	—	—	—	—	49
185—194	6	7	7	1	—	—	—	—	—	—	—	—	—	21
175—184	3	2	5	—	—	—	1	—	—	—	—	—	—	11
165—174	3	—	—	—	—	1	—	—	—	—	—	—	—	4
155—164	1	—	—	—	—	—	—	—	—	—	—	—	—	1
Totals	15	24	38	32	48	50	55	72	77	81	97	79	65	733

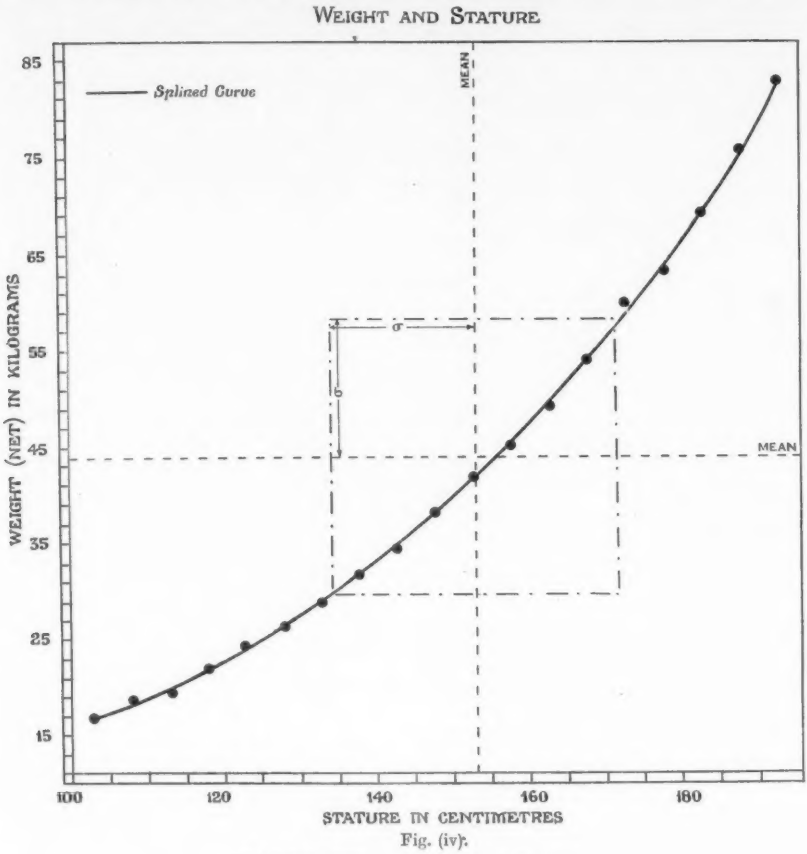


Fig. (iv).

TABLE VII. *Weight and Stature.*

Stature in Centimetres.

Weight (net) in Kilograms.	100·5—105·4	105·5—110·4	110·5—115·4	115·5—120·4	120·5—125·4	125·5—130·4	130·5—135·4	135·5—140·4	140·5—145·4	145·5—150·4	150·5—155·4	155·5—160·4	160·5—165·4	165·5—170·4	170·5—175·4	175·5—180·4	180·5—185·4	185·5—190·4	190·5—195·4	Totals
95·0—98·9	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	—	1
91·0—94·9	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0
87·0—90·9	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0
83·0—86·9	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	1	2
79·0—82·9	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	3	1	5
75·0—78·9	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	4	1	—	6
71·0—74·9	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	4	2	—	—	7
67·0—70·9	—	—	—	—	—	—	—	—	—	—	—	—	—	—	4	9	6	—	—	19
63·0—66·9	—	—	—	—	—	—	—	—	—	—	—	—	—	1	20	10	5	1	—	37
59·0—62·9	—	—	—	—	—	—	—	—	—	—	1	1	4	11	20	12	5	1	—	55
55·0—58·9	—	—	—	—	—	—	—	—	—	—	—	1	8	16	12	6	—	—	—	43
51·0—54·9	—	—	—	—	—	—	—	—	—	—	2	5	12	19	14	4	—	—	—	56
47·0—50·9	—	—	—	—	—	—	—	—	1	1	9	14	17	14	4	1	—	—	—	61
43·0—46·9	—	—	—	—	—	—	—	—	1	6	14	30	16	4	—	—	—	—	—	71
39·0—42·9	—	—	—	—	—	—	—	1	2	21	24	18	4	—	—	—	—	—	—	70
35·0—38·9	—	—	—	—	—	—	—	7	12	24	21	2	2	—	—	—	—	—	—	68
31·0—34·9	—	—	—	—	—	1	14	22	22	13	4	2	—	—	—	—	—	—	—	78
27·0—30·9	—	—	—	1	1	13	18	20	5	—	—	—	—	—	—	—	—	—	—	58
23·0—26·9	—	—	—	2	20	20	11	2	—	—	—	—	—	—	—	—	—	—	—	55
19·0—22·9	—	6	6	11	6	—	1	—	—	—	—	—	—	—	—	—	—	—	—	30
15·0—18·9	1	4	5	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	10
11·0—14·9	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1
Totals	1	11	11	14	27	34	44	52	43	65	75	73	63	65	75	47	24	7	2	733

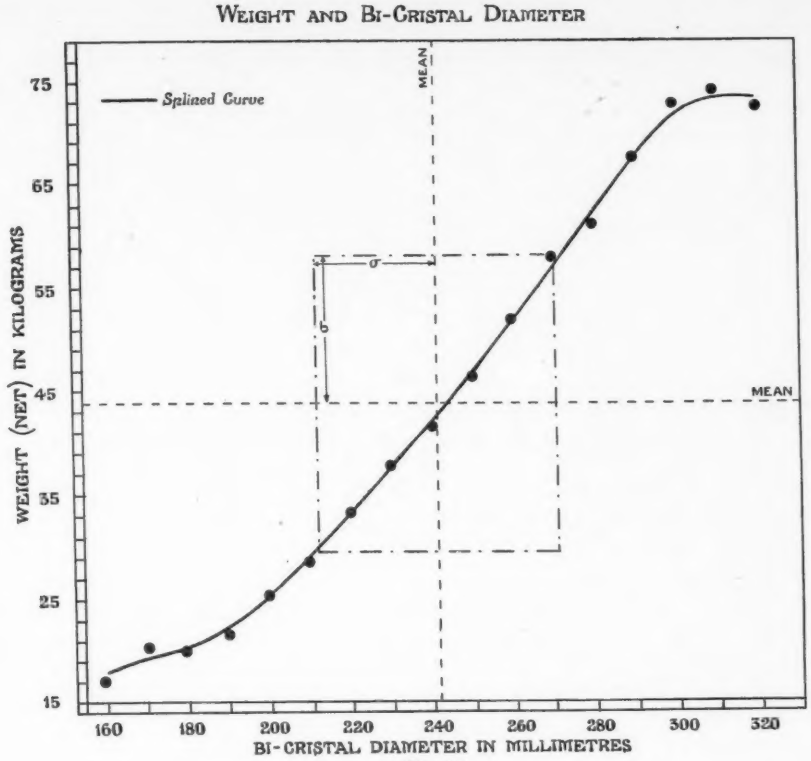


Fig. (v).

TABLE VIII. *Weight and Bi-cristal Diameter.*

Bi-cristal Diameter in Millimetres.

Weight (net) in Kilograms.	155—164	165—174	175—184	185—194	195—204	205—214	215—224	225—234	235—244	245—254	255—264	265—274	275—284	285—294	295—304	305—314	315—324	Totals
95·0—98·9	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	—	1
91·0—94·9	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0
87·0—90·9	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0
83·0—86·9	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	1	—	2
79·0—82·9	—	—	—	—	—	—	—	—	—	—	—	—	1	4	—	—	—	5
75·0—78·9	—	—	—	—	—	—	—	—	—	—	—	—	1	3	1	—	1	6
71·0—74·9	—	—	—	—	—	—	—	—	—	1	—	1	2	2	1	—	—	7
67·0—70·9	—	—	—	—	—	—	—	—	—	1	—	4	3	5	2	2	2	19
63·0—66·9	—	—	—	—	—	—	—	—	—	1	1	12	15	7	1	—	—	37
59·0—62·9	—	—	—	—	—	—	1	—	1	2	8	18	16	7	2	—	—	55
55·0—58·9	—	—	—	—	—	—	—	—	—	4	12	16	9	2	—	—	—	43
51·0—54·9	—	—	—	—	—	—	—	1	4	9	24	10	8	—	—	—	—	56
47·0—50·9	—	—	—	—	—	—	—	2	7	23	19	7	2	1	—	—	—	61
43·0—46·9	—	—	—	—	—	—	1	8	19	27	11	4	1	—	—	—	—	71
39·0—42·9	—	—	—	—	—	—	8	14	31	13	4	—	—	—	—	—	—	70
35·0—38·9	—	—	—	—	—	3	13	23	20	9	—	—	—	—	—	—	—	68
31·0—34·9	—	—	—	—	3	15	35	16	6	3	—	—	—	—	—	—	—	78
27·0—30·9	—	1	—	—	13	22	20	1	1	—	—	—	—	—	—	—	—	58
23·0—26·9	—	—	—	7	23	21	4	—	—	—	—	—	—	—	—	—	—	55
19·0—22·9	—	—	7	11	10	2	—	—	—	—	—	—	—	—	—	—	—	30
15·0—18·9	1	3	4	2	—	—	—	—	—	—	—	—	—	—	—	—	—	10
11·0—14·9	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	1
Totals	1	4	11	21	49	63	82	65	89	93	79	72	58	31	9	3	3	733

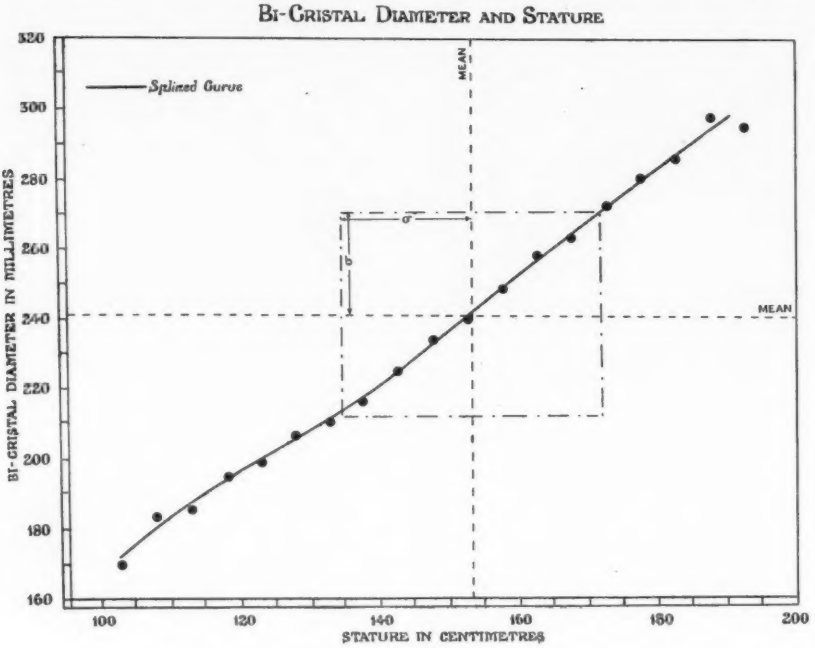


Fig. (vi).

TABLE IX.

Bi-cristal Diameter and Stature.

Stature in Centimetres.

Bi-cristal Diameter in Millimetres.	100·5—105·4	105·5—110·4	110·5—115·4	115·5—120·4	120·5—125·4	125·5—130·4	130·5—135·4	135·5—140·4	140·5—145·4	145·5—150·4	150·5—155·4	155·5—160·4	160·5—165·4	165·5—170·4	170·5—175·4	175·5—180·4	180·5—185·4	185·5—190·4	190·5—195·4	Totals
315—324	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	2	1	—	3
305—314	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	2	—	—	1	3
295—304	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	3	3	2	—	9
285—294	—	—	—	—	—	—	—	—	—	—	—	—	1	1	10	6	9	4	—	31
275—284	—	—	—	—	—	—	—	—	—	—	—	2	3	6	20	21	5	—	1	58
265—274	—	—	—	—	—	—	—	—	—	—	—	5	8	20	25	12	2	—	—	72
255—264	—	—	—	—	—	—	—	—	—	—	4	10	26	24	12	1	2	—	—	79
245—254	—	—	—	—	—	—	—	—	2	8	21	25	17	11	6	2	1	—	—	93
235—244	—	—	—	—	—	—	—	1	6	27	26	20	8	1	—	—	—	—	—	89
225—234	—	—	—	—	—	—	1	7	11	15	19	10	—	1	1	—	—	—	—	65
215—224	—	—	—	1	—	4	12	25	18	15	5	1	—	1	—	—	—	—	—	82
205—214	—	—	—	1	2	17	20	17	6	—	—	—	—	—	—	—	—	—	—	63
195—204	—	2	1	5	18	11	11	1	—	—	—	—	—	—	—	—	—	—	—	49
185—194	—	4	5	4	6	2	—	—	—	—	—	—	—	—	—	—	—	—	—	21
175—184	—	2	5	3	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	11
165—174	1	2	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	4
155—164	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1
Totals	1	11	11	14	27	34	44	52	43	65	75	73	63	65	75	47	24	7	2	733

CONCLUSIONS.

(1) The difference between the two multiple correlation coefficients is more than five times its probable error and therefore signifies that weight prediction from stature and the bi-cristal diameter of the pelvis is superior to prediction from stature and age*.

(2) The distributions, particularly those of y on x which are those which we wish to use, are not sufficiently curvilinear to invalidate conclusion (1) above, and therefore warrant at least provisional acceptance of that main inference.

Acknowledgment is due to Prof. L. L. Thurstone and Dr Luton Ackerson for advice on the statistical aspects, to Miss Margaret O'Connor for the computations and to Miss Marion Mason for preparing the diagrams.

REFERENCES.

(1) ROOT, H. F. and MILES, W. R.: "Physical Measurements of Diabetic Patients." *Journal of Metabolic Research*, Vol. II. p. 173, Aug. 1922, actually issued Jan. 1923.

(2) ROOT, H. F. and MILES, W. R.: "Physical Measurements on Operated Hyperthyroids." *Proceedings of the Society for Experimental Biology and Medicine*, Vol. XXIII. p. 727. May 1926.

(3) ROOT, H. F. and MILES, W. R.: "Weight and Physical Measurements after Thyroidectomy." *Archives of Internal Medicine*, Vol. XXXIX. p. 605, May 1927.

(4) GRAY, H. and PARMENTER, D. C.: "Chest Depth as an Index of Body Weight." *Journal of American Medical Association*, Vol. LXXXI. p. 2183, Dec. 29, 1923. This paper was based on correlation coefficients; it gives references to our earlier papers which date back to 1917.

(5) GRAY, H.: "Weight, Body Diameters and Age; Correlation Coefficients." *Proceedings of the Society for Experimental Biology and Medicine*, Vol. XXV. p. 384, Feb. 1928.

(6) HRDLIČKA, A.: *Anthropometry*. Philadelphia, 1920.

(7) BLAKEMAN, J., LEE, A. and PEARSON, K.: "A Study of the Biometric Constants of English Brain-Weights and their Relationships to External Physical Measurements." *Biometrika*, Vol. IV. p. 124, June 1905.

[* The probable errors of correlation coefficients of order ·93 to ·95 and therefore the probable error of their difference, assuming it to have been found from a correct formula, must be of very doubtful validity even in samples of so large a value as 800. The improvement of $R_{W.SBc}$ on $R_{W.SA}$ is so slight that it is difficult to say more than that *Bc* is as useful as *A*, which is a fact well worth noting. At the same time *R* for *curtailed* observations has rather a limited value, and the regression coefficients of weight, stature and bi-cristal diameter on age are of more importance than the correlation of these characters with age: see Appendix. ED.]

1905—1954	Totals
	3
1	3
	9
	31
1	58
	72
	79
	93
	89
	65
	82
	63
	49
	21
	11
	4
	1
2	733

APPENDIX.

ON THE CORRELATION OF TWO FIRST ORDER MULTIPLE CORRELATION COEFFICIENTS.

By K.P. and E.M.E.

The value of this correlation has already been given by Mr Philip Hall for the more general case*, and he has kindly provided us with its value in determinantal form for the case needed in a practical application, i.e. the correlation of $R_{1.23}$ with $R_{1.24}$.

We start with four variates 1, 2, 3, 4; a bar over a symbol denotes that the symbol has to be given its value in the sampled population, $R_{s.tuvw...}$ denotes a multiple correlation coefficient and $R_{st.uvw...}$ a partial correlation coefficient. Mr Hall has abbreviated these by supposing all the symbols of the universe of variates to be included and placing in brackets those which are supposed in a particular coefficient to be excluded. For example

$$R_{s\,(tu)} = R_{s.tvw...} \quad \text{and} \quad R_{st\,(v)} = R_{st.uw...}.$$

We shall deal with the correlation of $R^2_{1.23}$ and $R^2_{1.24}$ which will be the same as the correlation of $R_{1.23}$ and $R_{1.24}$, if the size of the sample be considerable.

Turning to Mr Hall's Equation (29) p. 106, we have the result needed if we insert the C-terms. These he has kindly provided; they are

$$\frac{D^2\left(1\,\frac{3}{4}\right)}{D(13)\,D(14)} + 1 + \frac{D^2_{34}}{D(3)\,D(4)} - \frac{D\left(3\,\frac{1}{4}\right)D_{14} + D\left(1\,\frac{3}{4}\right)D_{34}}{D(13)\,D(4)} - \frac{D\left(4\,\frac{1}{3}\right)D_{13} + D\left(1\,\frac{3}{4}\right)D_{34}}{D(14)\,D(3)},$$

and thus lead to the result:

$$R^2_{1\,23}\,R^2_{1.24} = \frac{1}{2\bar{R}_{1.23}\,\bar{R}_{1.24}}\left[\bar{R}^2_{1.23} + \bar{R}^2_{1.24} + \bar{R}^2_{34.12} + \bar{R}^2_{34.2} - \bar{R}_{14.2}\,\bar{R}_{14.23}\sqrt{\frac{D(43)\,D(1)}{D(13)\,D(4)}} - \bar{R}_{13.2}\,\bar{R}_{13.24}\sqrt{\frac{D(43)\,D(1)}{D(14)\,D(3)}} - \bar{R}_{34.2}\,\bar{R}_{34.12}\left\{\sqrt{\frac{D(14)\,D(3)}{D(13)\,D(4)}} + \sqrt{\frac{D(13)\,D(4)}{D(14)\,D(3)}}\right\}\right].$$

* *Biometrika*, Vol. XIX. pp. 100—109.

But since $\dfrac{D(\beta)}{D(\gamma\beta)} = 1 - \overline{R}^2\gamma(\beta)$ we can replace the determinants by multiple correlation coefficients, and we find:

$$r_{R^2_{1.24}\,R^2_{1.23}} = \frac{1}{2\overline{R}_{1.23}\,\overline{R}_{1.24}}\left[\overline{R}^2_{1.23} + \overline{R}^2_{1.24} + \overline{R}^2_{34.12} + \overline{R}^2_{34.2}\right.$$
$$- \overline{R}_{14.2}\,\overline{R}_{14.23}\sqrt{\frac{1-\overline{R}^2_{3.24}}{1-\overline{R}^2_{3.12}}} - \overline{R}_{13.2}\,\overline{R}_{13.24}\sqrt{\frac{1-\overline{R}^2_{4.23}}{1-\overline{R}^2_{4.12}}}$$
$$\left. - \overline{R}_{34.2}\,\overline{R}_{34.12}\left\{\sqrt{\frac{1-\overline{R}^2_{1.24}}{1-\overline{R}^2_{1.23}}} + \sqrt{\frac{1-\overline{R}^2_{1.23}}{1-\overline{R}^2_{1.24}}}\right\}\right] \quad \text{.....(i).}$$

To test this formula we suppressed the variate 2, and after a fair amount of algebra found it reduced to

$$r_{r^2_{14}\,r^2_{13}} = r_{34} - \frac{1}{2}\frac{r_{14}\,r_{13}\,(1 - r^2_{14} - r^2_{13} - r^2_{34} + 2r_{14}r_{13}r_{34}}{(1-r^2_{13})(1-r^2_{14})} \quad \text{.....(ii),}$$

thus agreeing to our order of approximation with the Pearson-Filon formula* for $r_{r_{14}r_{13}}$, as it should do.

It is possible that using the known relations between multiples and partials, the above formula could be reduced to a simpler form, but we have not succeeded in doing so, and we see so far no reason for supposing that the Pearson-Filon formula can be directly applied to multiple coefficients, as some have suggested.

In Dr Gray's paper we have for the variates:

1 = Weight, 2 = Stature, 3 = Bicristal diameter, 4 = Age;

and he finds

$R_{1.23} = \cdot 9411$, $R_{1.24} = \cdot 9314$ or $R^2_{1.23} = \cdot 8856{,}6921$ and $R^2_{1.24} = \cdot 8675{,}0596$,

and the problem is whether the difference between these last two is, or is not, significant.

As usual, we do not know values in the sampled population, and have to use R for $\overline{R}$ throughout. Further, since all Dr Gray's age correlation tables are doubly curtailed, although the curtailing of the upper age limit is the more serious, the total correlation coefficients, and with them the multiples, are deprived of much real significance. It is clearly the regression, not the correlation coefficients which will have real significance†. However, to illustrate the use of the above formula we may ask the question: How does the difference $R^2_{1.23} - R^2_{1.24}$ stand in relation to its "probable error"? We have

$$\sigma^2_{R^2_{1.23}} = \frac{4R^2_{1.23}(1 - R^2_{1.23})^2}{N} = \cdot 000062176,$$

$$\sigma^2_{R^2_{1.24}} = \frac{4R^2_{1.24}(1 - R^2_{1.24})^2}{N} = \cdot 000083104,$$

since $N = 733$.

* *Phil. Trans.* Vol. 191 A, p. 259. Eqn (xxxvii). 1898.

† There is also another grave difficulty, even with the relatively large numbers that are involved namely, what may be the real significance of the probable error of the difference of correlations, total or multiple, when they are of such high values?

But the probable error of the difference $R^2_{1.23} - R^2_{1.24}$

$$= \cdot 67449\sqrt{\sigma^2_{R^2_{1.23}} + \sigma^2_{R^2_{1.24}} - 2\sigma_{R^2_{1.23}}\sigma_{R^2_{1.24}}r_{R^2_{1.23}}r_{R^2_{1.24}}} \quad \ldots\ldots\ldots\text{(iii)}.$$

Now we find from (ii)

$$r_{R^2_{1.23}R^2_{1.24}} = \cdot 921{,}672,$$

which gives:

Probable Error of Difference $R^2_{1.23} - R^2_{1.24}$, or ·01816, is ·002406 by formula (iii), or the difference is about 7·5 times its probable error.

If we take $r_{R_{1.23}R_{1.24}} = r_{R^2_{1.23}R^2_{1.24}}$, we have for probable error of difference $R_{1.23} - R_{1.24}$ the expression

$$\cdot 67449\sqrt{\sigma^2_{R_{1.23}} + \sigma^2_{R_{1.24}} - 2\sigma_{R_{1.23}}\sigma_{R_{1.24}}r_{R_{1.23}R_{1.24}}},$$

where $$\sigma_{R_{1.23}} = \frac{1 - R^2_{1.23}}{\sqrt{N}} \quad \text{and} \quad \sigma_{R_{1.24}} = \frac{1 - R^2_{1.24}}{\sqrt{N}},$$

i.e. Probable Error of Difference $R_{1.23} - R_{1.24}$, or ·0097, equals ·0013, or the difference is again about 7·5 times its probable error.

Accordingly, if we could trust the theory here applied, the prediction of Weight from Age and Bicristal diameter would be significantly better than from Age and Stature. The corresponding formulae are as follows:

Multiple Regression Formula of Weight on Bicristal Diameter and Age:

Probable Weight in kilos

$$= \cdot 3144\,(\text{Bicristal in mm.}) + 1\cdot 3954\,(\text{Age in years}) - 49\cdot 3803 \pm 3\cdot 678/\sqrt{n} \quad \ldots\text{(iv)}.$$

Multiple Regression Formula of Weight on Stature and Age:

Probable Weight in kilos

$$= \cdot 7010\,(\text{Stature in cms.}) + \cdot 1231\,(\text{Age in years}) - 63\cdot 942 \pm 3\cdot 551/\sqrt{n} \quad \ldots\text{(v)}.$$

Multiple Regression Formula of Weight on Bicristal Diameter and Stature:

Probable Weight in kilos

$$= \cdot 1768\,(\text{Bicristal in mm.}) + \cdot 4637\,(\text{Stature in cms.}) - 69\cdot 863 \pm 3\cdot 299/\sqrt{n} \quad \ldots\text{(vi)},$$

where n is the number of individuals for which the probable Weight is determined.

These equations indicate at once that, if we use only a bivariate prediction formula, the prediction from Stature and Bicristal diameter (vi) will give the best result. Age is of less importance, although it is better to use Stature and Age to find Weight than Bicristal and Age. However, all three prediction formulae have a difference in their probable errors of the order ·40 kilos only, or one of less than 1 °/₀ of the mean weight. Thus in an individual case there is little to choose between them.

It appeared worth while testing whether we could obtain increased accuracy of prediction by including Age as a determining factor as well as Stature and Bicristal diameter. The corresponding triple regression formula is:

Probable Weight in kilos

$$= \cdot 1768\,(\text{Bicristal in mms.}) + \cdot 4659\,(\text{Stature in cms.}) + \cdot 0066\,(\text{Age in years}) - 70\cdot 1332 \pm 3\cdot 299/\sqrt{n} \quad \ldots\text{(vii)}.$$

It will be seen on comparing with Equation (vi) that we have gained absolutely nothing in our prediction formula or its probable error by adding Age; the greatest influence of Age cannot exceed about 1/10 of a kilo*. Indeed, the multiple correlation coefficient of Weight on Stature, Bicristal Diameter and Age $R_{1.234} = \cdot 9411$, which is precisely the value found for $R_{1.23}$, or the multiple correlation coefficient of Weight on Stature and Bicristal Diameter. This is another illustration of the familiar fact that to increase the number of very highly inter-correlated variates will not sensibly increase the accuracy of a prediction formula.

Reminding the reader that of our subscripts 1 = Weight, 2 = Stature, 3 = Bicristal Diameter and 4 = Age, we may add some additional multiple and partial correlation coefficients which have been computed from Dr Gray's data in the course of our work:

$$\begin{array}{lll} R_{1.234} = \cdot 9411, & [R_{1.23} = \cdot 9411, & R_{1.24} = \cdot 9314]\dagger, \\ R_{4.23} = \cdot 9342, & R_{4.12} = \cdot 9339, & R_{1.34} = \cdot 9262, \\ R_{3.24} = \cdot 9267, & R_{3.12} = \cdot 9367, & \\ R_{34.12} = \cdot 0723, & R_{14.23} = -\cdot 0016, & R_{13.24} = \cdot 3707, \\ R_{34.2} = \cdot 0772, & R_{14.2} = \cdot 0273, & R_{13.2} = \cdot 3716. \end{array}$$

Several of these appear to have very suggestive values‡, which we do not stay now to discuss. The reader who does consider them should, however, bear in mind that when the Age, variate 4, appears we are dealing with curtailed distributions, and the corresponding coefficient must accordingly be interpreted as having reference solely to distributions curtailed in precisely the same manner. The regression equations show us that within this prescribed range the Bicristal Diameter has greater value than Age for Weight prediction.

* The clinicians who are seeking the "normal weight" of an individual are, we believe, following a Will o' the Wisp. They cannot get rid of wide variabilities by any prediction formula. Here, with very high correlations, the variability is still 3·3 kilos, or about 7 lbs. Luckily it is so, or, for the purposes of selective evolution, there would be small chance of modifying one character without upsetting the whole organism.

† Given by Dr Gray.

‡ Which differ in a few cases rather widely from the values in Table II of Dr Gray's paper; the latter were, however, determined for a somewhat different population.

ON CORRECTIONS FOR THE MOMENT-COEFFICIENTS OF FREQUENCY DISTRIBUTIONS WHEN THERE ARE INFINITE ORDINATES AT ONE OR BOTH OF THE TERMINALS OF THE RANGE.

By GERTRUDE E. PEARSE, M.A. (Cantab.).

(1) A PAPER* by Eleanor Pairman and Karl Pearson, F.R.S., was published in 1918 on "Corrections for the Moment-Coefficients of Limited Range Frequency Distributions when there are Finite or Infinite Ordinates and any Slopes at the Terminals of the Range." In the second part of that paper, dealing with cases of asymptotic frequency, the auxiliary curve selected to give the first five frequencies was

$$Z = \int_x^{x_p} y\,dx = N\,[1 + x^q\,(A + Bx + Cx^2 + Dx^3 + Ex^4)].$$

It was remarked that to obtain an infinite ordinate and finite area, q could be given any value from slightly over zero to slightly under unity, and that the size of q would measure so to speak the intensity of the asymptoting. Since, however, the determination of q would involve great labour, it was given the mean value one half.

To test the accuracy of the results obtained by this method, let us select the curve $y = \dfrac{\cdot 2^{\cdot 1} e^{-\cdot 2x} x^{-\cdot 9}}{\Gamma(\cdot 1)}$, which can be expanded in the form of our auxiliary curve with $q = \cdot 1$. Now

$$\int_0^x y\,dx = \frac{1}{\Gamma(\cdot 1)} \int_0^{\cdot 2x} e^{-v} v^{-\cdot 9}\,dv = I\,(u, p), \quad \text{if} \quad p = -\cdot 9, \quad \text{and} \quad u = \frac{v}{\sqrt{\cdot 1}} = \frac{\cdot 2x}{\sqrt{\cdot 1}},$$

where I is the Incomplete Γ-function, the value of which can be found from the tables. The frequencies are given in the Table on p. 315.

Taking moments about $x = \cdot 5\sqrt{\cdot 1}$,

$$\cdot 169{,}9490\nu_1'' = \cdot 246{,}685{,}35\sqrt{\cdot 1}.$$

From equations (xxxii),

$$\cdot 169{,}9490a_1' = -\cdot 108{,}6639,$$
$$\cdot 169{,}9490a_3' = -\cdot 010{,}7457,$$
$$\cdot 169{,}9490a_5' = -\cdot 001{,}9342.$$

Therefore

$$\cdot 169{,}9490\,\frac{h}{12}\,(a_1' - \tfrac{1}{60}a_3' + \tfrac{1}{2520}a_5') = -\cdot 004{,}1605\sqrt{\cdot 1}.$$

From equations (xxxi),

$$\cdot 742{,}0263\mu_1'' = -\cdot 291{,}2096\sqrt{\cdot 1}.$$

Thus

$$\cdot 911{,}9753\mu_1' = -\cdot 048{,}6848\sqrt{\cdot 1}.$$

* *Biometrika*, Vol. XII. pp. 231 *et seq.*

On transferring the origin to the start of the curve, $\cdot 911{,}9753\mu_1' = \cdot 128{,}800$, whereas the true value of the first moment obtained by integration is $\cdot 114{,}564$, an error of more than 12 %. The crude, uncorrected value is $\cdot 163{,}541$. Our result is a great improvement on this, but is scarcely as good as we should like. This is not surprising, since we have taken $q = \cdot 5$ when it should be $\cdot 1$.

x	u	$I(u, p)$	Frequencies	Distances of Centres from $x = \cdot 5\sqrt{\cdot 1}$
0	0	0		
			$\cdot 742{,}0263$	
$0 \cdot 5\sqrt{\cdot 1}$	$\cdot 1$	$\cdot 742{,}0263$		
			$\cdot 051{,}0210$	$\frac{\sqrt{\cdot 1}}{4}$
$1 \cdot 0\sqrt{\cdot 1}$	$\cdot 2$	$\cdot 793{,}0473$		
			$\cdot 030{,}5247$	$3\frac{\sqrt{\cdot 1}}{4}$
$1 \cdot 5\sqrt{\cdot 1}$	$\cdot 3$	$\cdot 823{,}5720$		
			$\cdot 021{,}7166$	$5\frac{\sqrt{\cdot 1}}{4}$
$2 \cdot 0\sqrt{\cdot 1}$	$\cdot 4$	$\cdot 845{,}2886$		
			$\cdot 016{,}7398$	$7\frac{\sqrt{\cdot 1}}{4}$
$2 \cdot 5\sqrt{\cdot 1}$	$\cdot 5$	$\cdot 862{,}0284$		
			$\cdot 013{,}5217$	$9\frac{\sqrt{\cdot 1}}{4}$
$3 \cdot 0\sqrt{\cdot 1}$	$\cdot 6$	$\cdot 875{,}5501$		
			$\cdot 011{,}2637$	$11\frac{\sqrt{\cdot 1}}{4}$
$3 \cdot 5\sqrt{\cdot 1}$	$\cdot 7$	$\cdot 886{,}8138$		
			$\cdot 009{,}5898$	$13\frac{\sqrt{\cdot 1}}{4}$
$4 \cdot 0\sqrt{\cdot 1}$	$\cdot 8$	$\cdot 896{,}4036$		
			$\cdot 008{,}2988$	$15\frac{\sqrt{\cdot 1}}{4}$
$4 \cdot 5\sqrt{\cdot 1}$	$\cdot 9$	$\cdot 904{,}7024$		
			$\cdot 007{,}2729$	$17\frac{\sqrt{\cdot 1}}{4}$
$5 \cdot 0\sqrt{\cdot 1}$	$1 \cdot 0$	$\cdot 911{,}9753$		

(2) In order to better our approximation let us proceed, therefore, to calculate values of the abruptness and moment coefficients for different values of q between 0 and 1.

If Nn_1', Nn_2', Nn_3', Nn_4', Nn_5' be the first five frequencies, and the base interval h be unity, then since

$$Z = N[1 + x^q(A + Bx + Cx^2 + Dx^3 + Ex^4)],$$

$$\left.\begin{aligned} A + B + C + D + E &= -n_1' \\ 2^q(A + 2B + 2^2C + 2^3D + 2^4E) &= -n_1' - n_2' \\ 3^q(A + 3B + 3^2C + 3^3D + 3^4E) &= -n_1' - n_2' - n_3' \\ 4^q(A + 4B + 4^2C + 4^3D + 4^4E) &= -n_1' - n_2' - n_3' - n_4' \\ 5^q(A + 5B + 5^2C + 5^3D + 5^4E) &= -n_1' - n_2' - n_3' - n_4' - n_5' \end{aligned}\right\} \quad \ldots\ldots(\mathrm{I}).$$

These equations on solution give

$$\left.\begin{aligned}
A &= n_1'\left(-5+\frac{10}{2^q}-\frac{10}{3^q}+\frac{5}{4^q}-\frac{1}{5^q}\right)+n_2'\left(\frac{10}{2^q}-\frac{10}{3^q}+\frac{5}{4^q}-\frac{1}{5^q}\right)\\
&\quad+n_3'\left(-\frac{10}{3^q}+\frac{5}{4^q}-\frac{1}{5^q}\right)+n_4'\left(\frac{5}{4^q}-\frac{1}{5^q}\right)+n_5'\left(-\frac{1}{5^q}\right)\\
12B &= n_1'\left(77-\frac{214}{2^q}+\frac{234}{3^q}-\frac{122}{4^q}+\frac{25}{5^q}\right)+n_2'\left(-\frac{214}{2^q}+\frac{234}{3^q}-\frac{122}{4^q}+\frac{25}{5^q}\right)\\
&\quad+n_3'\left(\frac{234}{3^q}-\frac{122}{4^q}+\frac{25}{5^q}\right)+n_4'\left(-\frac{122}{4^q}+\frac{25}{5^q}\right)+n_5'\left(\frac{25}{5^q}\right)\\
24C &= n_1'\left(-71+\frac{236}{2^q}-\frac{294}{3^q}+\frac{164}{4^q}-\frac{35}{5^q}\right)+n_2'\left(\frac{236}{2^q}-\frac{294}{3^q}+\frac{164}{4^q}-\frac{35}{5^q}\right)\\
&\quad+n_3'\left(-\frac{294}{3^q}+\frac{164}{4^q}-\frac{35}{5^q}\right)+n_4'\left(\frac{164}{4^q}-\frac{35}{5^q}\right)+n_5'\left(-\frac{35}{5^q}\right)\\
12D &= n_1'\left(7-\frac{26}{2^q}+\frac{36}{3^q}-\frac{22}{4^q}+\frac{5}{5^q}\right)+n_2'\left(-\frac{26}{2^q}+\frac{36}{3^q}-\frac{22}{4^q}+\frac{5}{5^q}\right)\\
&\quad+n_3'\left(\frac{36}{3^q}-\frac{22}{4^q}+\frac{5}{5^q}\right)+n_4'\left(-\frac{22}{4^q}+\frac{5}{5^q}\right)+n_5'\left(\frac{5}{5^q}\right)\\
24E &= n_1'\left(-1+\frac{4}{2^q}-\frac{6}{3^q}+\frac{4}{4^q}-\frac{1}{5^q}\right)+n_2'\left(\frac{4}{2^q}-\frac{6}{3^q}+\frac{4}{4^q}-\frac{1}{5^q}\right)\\
&\quad+n_3'\left(-\frac{6}{3^q}+\frac{4}{4^q}-\frac{1}{5^q}\right)+n_4'\left(\frac{4}{4^q}-\frac{1}{5^q}\right)+n_5'\left(-\frac{1}{5^q}\right)
\end{aligned}\right\}\quad\text{...(II).}$$

The moments of the first frequency about $x=1$ can be calculated from the integral $n_1\mu_s''=\frac{1}{N}\int_0^1(1-x)^q\frac{dZ}{dx}\,dx$, whence after a little reduction we obtain

$$\left.\begin{aligned}
n_1\mu_1' &= \frac{A}{q+1}+\frac{B}{q+2}+\frac{C}{q+3}+\frac{D}{q+4}+\frac{E}{q+5}\\
n_1\mu_2'' &= -2\left\{\frac{A}{(q+1)(q+2)}+\frac{B}{(q+2)(q+3)}+\frac{C}{(q+3)(q+4)}+\frac{D}{(q+4)(q+5)}\right.\\
&\qquad\left.+\frac{E}{(q+5)(q+6)}\right\}\\
n_1\mu_3'' &= 6\left\{\frac{A}{(q+1)(q+2)(q+3)}+\frac{B}{(q+2)(q+3)(q+4)}+\frac{C}{(q+3)(q+4)(q+5)}\right.\\
&\qquad\left.+\frac{D}{(q+4)(q+5)(q+6)}+\frac{E}{(q+5)(q+6)(q+7)}\right\}\\
n_1\mu_4'' &= -24\left\{\frac{A}{(q+1)(q+2)(q+3)(q+4)}+\frac{B}{(q+2)(q+3)(q+4)(q+5)}\right.\\
&\qquad+\frac{C}{(q+3)(q+4)(q+5)(q+6)}+\frac{D}{(q+4)(q+5)(q+6)(q+7)}\\
&\qquad\left.+\frac{E}{(q+5)(q+6)(q+7)(q+8)}\right\}
\end{aligned}\right\}\quad\text{(III).}$$

Substituting the values obtained for A, B, C, D and E, we obtain

$$\begin{aligned}\mu_s'' = \Big\{ & n_1' \left(\nu_s' + \frac{\nu_s''}{2^q} + \frac{\nu_s'''}{3^q} + \frac{\nu_s''''}{4^q} + \frac{\nu_s'''''}{5^q} \right) \\ & + n_2' \left(\frac{\nu_s''}{2^q} + \frac{\nu_s'''}{3^q} + \frac{\nu_s''''}{4^q} + \frac{\nu_s'''''}{5^q} \right) + n_3' \left(\frac{\nu_s'''}{3^q} + \frac{\nu_s''''}{4^q} + \frac{\nu_s'''''}{5^q} \right) \\ & + n_4' \left(\frac{\nu_s''''}{4^q} + \frac{\nu_s'''''}{5^q} \right) + n_5' \left(\frac{\nu_s'''''}{5^q} \right) \Big\}\end{aligned}$$

where

$$\nu_1 = \frac{w_1}{24(q+1)(q+2)(q+3)(q+4)(q+5)}$$

$$\nu_2 = \frac{w_2}{12(q+1)(q+2)\dots(q+6)}$$

$$\nu_3 = \frac{w_3}{4(q+1)(q+2)\dots(q+7)}, \qquad \nu_4 = \frac{w_4}{(q+1)(q+2)\dots(q+8)}$$

$$\begin{aligned}
w_1''''' &= -2(3q^3+47q^2+258q+502) \\
w_1'''' &= 8(4q^3+62q^2+335q+637) \\
w_1''' &= -12(6q^3+91q^2+477q+872) \\
w_1'' &= 8(12q^3+170q^2+825q+1387) \\
w_1' &= -24q^4-386q^3-2374q^2-6736q-7604 \\
w_2''''' &= 12q^3+246q^2+1758q+4404 \\
w_2'''' &= -64q^3-1296q^2-9104q-22272 \\
w_2''' &= 144q^3+2844q^2+19332q+45432 \\
w_2'' &= -192q^3-3504q^2-21936q-47424 \\
w_2' &= 24q^4+532q^3+4566q^2+18158q+28500 \\
w_3''''' &= -18q^3-456q^2-4002q-12204 \\
w_3'''' &= 96q^3+2400q^2+20688q+61584 \\
w_3''' &= -216q^3-5256q^2-43776q-125136 \\
w_3'' &= 288q^3+6432q^2+49200q+129456 \\
w_3' &= -24q^4-678q^3-7416q^2-37422q-73860 \\
w_4''''' &= -24q^3-724q^2-7524q-26984 \\
w_4'''' &= 128q^3+3808q^2+38848q+135968 \\
w_4''' &= -288q^3-8328q^2-82008q-275568 \\
w_4'' &= 384q^3+10144q^2+91584q+283424 \\
w_4' &= -24q^4-824q^3-10924q^2-66484q-157160
\end{aligned} \qquad \text{(IV).}$$

From these equations Table II was calculated giving the first four moments of the first frequency in terms of the relative frequencies n_1', n_2', n_3', n_4', n_5'.

The abruptness coefficients at the point $x=1$ can then be found from the formula

$$a_s = \frac{1}{N} \left(\frac{d^s Z}{dx^s} \right)_{x=1},$$

whence

$$\left.\begin{aligned}
a_1 &= Aq + B(q+1) + C(q+2) + D(q+3) + E(q+4)\\
a_2 &= Aq(q-1) + Bq(q+1) + C(q+1)(q+2) + D(q+2)(q+3)\\
&\qquad + E(q+3)(q+4)\\
a_3 &= Aq(q-1)(q-2) + Bq(q+1)(q-1) + Cq(q+1)(q+2)\\
&\qquad + D(q+1)(q+2)(q+3) + E(q+2)(q+3)(q+4)\\
a_4 &= A(q-3)(q-2)(q-1)q + B(q-2)(q-1)q(q+1)\\
&\qquad + C(q-1)q(q+1)(q+2) + Dq(q+1)(q+2)(q+3)\\
&\qquad + E(q+1)(q+2)(q+3)(q+4)\\
a_5 &= A(q-4)(q-3)(q-2)(q-1)q + B(q-3)(q-2)(q-1)q(q+1)\\
&\qquad + C(q-2)(q-1)q(q+1)(q+2) + D(q-1)q(q+1)(q+2)(q+3)\\
&\qquad + E(q)(q+1)(q+2)(q+3)(q+4)
\end{aligned}\right\} \text{(V).}$$

Substituting for A, B, C, D and E we obtain

$$\left.\begin{aligned}
a_s = n_1'&\left[u_s' + \frac{u_s''}{2^q} + \frac{u_s'''}{3^q} + \frac{u_s''''}{4^q} + \frac{u_s'''''}{5^q}\right] + n_2'\left[\frac{u_s''}{2^q} + \frac{u_s'''}{3^q} + \frac{u_s''''}{4^q} + \frac{u_s'''''}{5^q}\right]\\
&+ n_3'\left[\frac{u_s'''}{3^q} + \frac{u_s''''}{4^q} + \frac{u_s'''''}{5^q}\right] + n_4'\left[\frac{u_s''''}{4^q} + \frac{u_s'''''}{5^q}\right] + n_5'\left[\frac{u_s'''''}{5^q}\right]
\end{aligned}\right\}$$

where

$$\left.\begin{aligned}
&u_2' = -\tfrac{1}{12}(q^2 - 62q + 35), \quad u_2'' = -\tfrac{1}{3}(24q - 26), \quad u_2''' = \tfrac{1}{2}(12q - 19)\\
&u_2'''' = \tfrac{1}{3}(-8q + 14), \quad u_2''''' = \tfrac{1}{12}(6q - 11), \quad u_3' = \tfrac{1}{4}(-4q^3 + 37q^2 - 68q + 10)\\
&u_3'' = 12q^2 + 38q - 9, \quad u_3''' = \tfrac{1}{2}(18q^2 - 75q + 24), \quad u_3'''' = -4q^2 + 18q - 7\\
&u_3''''' = \tfrac{1}{4}(3q^2 - 14q + 6), \quad u_4' = \tfrac{1}{6}(-6q^4 + 86q^3 - 321q^2 + 301q - 6)\\
&u_4'' = 4(-4q^3 + 33q^2 - 50q + 3), \quad u_4''' = 3(4q^3 - 31q^2 + 43q - 2)\\
&u_4'''' = \tfrac{4}{3}(-4q^3 + 33q^2 - 50q + 3), \quad u_4''''' = \tfrac{1}{2}(2q^3 + 17q^2 + 27q - 2)\\
&u_5' = \tfrac{1}{24}(-24q^5 + 490q^4 - 3040q^3 + 6650q^2 - 4196q)\\
&u_5'' = \frac{10q}{3}(-6q^3 + 62q^2 - 171q + 121), \quad u_5''' = 5q(3q^3 - 37q^2 + 114q - 86)\\
&u_5'''' = \frac{10q}{3}(-2q^3 + 26q^2 - 85q + 67), \quad u_5''''' = \frac{10q}{24}(3q^3 - 40q^2 + 135q - 110)
\end{aligned}\right\} \text{(VI),}$$

with similar expressions for u_1.

From these equations Table III was calculated, giving the abruptness coefficients at $x = 1$ in terms of the first five relative frequencies.

Now let a_s be the value of a_s' where the true frequencies are substituted for the relative frequencies, and let

$$\left.\begin{aligned}
K_1 &= n_1\mu_1'' + \frac{a_1}{12} - \frac{a_3}{720} + \frac{a_5}{12.2520}\\
K_2 &= n_1\mu_2'' - \frac{a_2}{120} + \frac{1}{24.126}a_4\\
K_3 &= n_1\mu_3'' - \frac{a_1}{40} + \frac{a_3}{504} - \frac{a_5}{9600}\\
K_4 &= n_1\mu_4'' + \frac{a_2}{126} - \frac{a_4}{1440}
\end{aligned}\right\} \text{.....(VII).}$$

Then the corrected moments about the end of the first subrange are given by

$$\left.\begin{aligned} N\mu_1'' &= N_{2p}\nu_1'' + hK_1 \\ N\mu_2'' &= N_{2p}(\nu_2'' - \tfrac{1}{12}h^2) + h^2K_2 \\ N\mu_3'' &= N_{2p}(\nu_3'' - \tfrac{1}{4}h^2\nu_1'') + h^3K_3 \\ N\mu_4'' &= N_{2p}(\nu_4'' - \tfrac{1}{2}h^2\nu_2'' + \tfrac{7}{240}h^4) + h^4K_4 \end{aligned}\right\} \quad \text{.....(VIII),}$$

where $N_{2p} = N - n_1$.

The values of the K's as calculated from equations (VII), are given in Table IV in terms of the first five true frequencies. All the computer has to do, therefore, to obtain the corrected moments is to calculate the K's from Table IV and add them to the crude moments about the end of the first subrange, corrected by Shepherd's corrections. The transference to the mean then proceeds in the usual manner. Tables II and III will not usually be required.

In the case of both the moments and the abruptness coefficients it will be found more satisfactory to substitute for A, B, C, D and E as above, rather than to table these constants first and then deduce Tables III and IV from them, as, in the latter method, such a large number of figures have to be retained in the constants in order to obtain even a few decimal places in the moments and abruptness coefficients.

Let us now calculate the mean and standard deviation of our original example $y = \dfrac{\cdot 2^{\cdot 1} e^{-\cdot 2x} x^{-\cdot 9}}{\Gamma(\cdot 1)}$ from our tables, giving q the correct value, namely ·1.

We have for the crude moments about $x = h$ ($h = \cdot 5\sqrt{\cdot 1}$),

$N_{2p}\nu_1'' = \cdot 246{,}685{,}35\sqrt{\cdot 1}$ and $N_{2p}\nu_2'' = \cdot 060{,}8553$ and $\frac{1}{12}N_{2p}h^2 = \cdot 000{,}3541$.

Table IV gives

$$K_1 = -\cdot 681{,}364{,}068 \quad \text{and} \quad K_2 = \cdot 643{,}192{,}136.$$

Therefore from equations (VIII) corrected moments about $x = h$ are

$$N\mu_1'' = -\cdot 093{,}996{,}68\sqrt{\cdot 1} \quad \text{and} \quad N\mu_2'' = \cdot 076{,}5810.$$

Whence Mean = ·114,470 and (S.D.)^2 = ·082,9103.

The true values are

Mean = ·114,564 and (S.D.)^2 = ·083,196,70.

The errors are therefore ·08 °/₀ and ·34 °/₀, whereas with $q = \cdot 5$ the error in the mean was 12 °/₀. Thus even when more than ·8 of the total frequency is contained in the first subrange our tables give very reasonable results; and we see that the correct choice of q makes a very considerable difference in our curve constants.

(3) In the next place let us turn our attention to the problem of selecting q, first supposing that we know the values of the frequencies on equal subranges only. Several different methods are possible, of which the simplest seems to be to choose q to give the best fit to the sixth subrange. The five constants A, B, C, D and E have already been chosen so that our auxiliary curve fits the first five sub-frequencies for all values of q. Now although the auxiliary curve is not used for

any calculations beyond the first subrange and the discovery of the abruptness coefficients at its end, we shall still ensure greater continuity if we make it fit as nearly as possible to the sixth subrange. This method, though possessing certain disadvantages, has one point very much in its favour, namely, that it seems to be the simplest for the computer. It might therefore well be used as a first attempt. If, however, for some reason, such as large errors in the data near the asymptote, a bad fit is obtained, there are other methods which might be tried.

To select q by this means we have the equation

$$6^q(A+6B+6^2C+6^3D+6^4E)=-(n_1'+n_2'+n_3'+n_4'+n_5'+n_6'),$$

which gives n_6' in terms of the first five frequencies. These values are given in Table I. In order to select q in any given case, we therefore calculate the value of n_6' from Table I for one or two values of q and select that value of q which gives the closest approximation to the given value of n_6'. If the asymptoting seems to be extreme we naturally first try low values of q and vice versa. In practice this is not very tedious since three decimal places will usually be sufficient.

For instance, consider the curve

$$y=\frac{1}{\Gamma(\cdot5)}e^{-x}x^{-\cdot5},$$

taking h as $\sqrt{\cdot5}$. $\int_0^x y\,dx=I(u,p)$, where $p=-\cdot5$, $u=\dfrac{x}{\sqrt{\cdot5}}$ and I is the Incomplete Γ-function. We have from the *Tables of the Incomplete* Γ*-function*,

x	u	$I(u,p)$	Frequencies
0	0	0	
			$\cdot765,6418=n_1'$
$\sqrt{\cdot5}$	1	$\cdot765,6418$	
			$\cdot141,7492=n_2'$
$2\sqrt{\cdot5}$	2	$\cdot907,3910$	
			$\cdot053,1882=n_3'$
$3\sqrt{\cdot5}$	3	$\cdot960,5792$	
			$\cdot022,0335=n_4'$
$4\sqrt{\cdot5}$	4	$\cdot982,6127$	
			$\cdot009,5534=n_5'$
$5\sqrt{\cdot5}$	5	$\cdot992,1661$	
			$\cdot004,2536=n_6'$
$6\sqrt{\cdot5}$	6	$\cdot996,4197$	

On testing $q=\cdot8$ in Table I we obtain $n_6'=\ \cdot165$,

$q=\cdot4$ $\qquad\qquad$ $n_6'=-\cdot019$.

The best value of q lies therefore between these and is nearer to $\cdot4$. Trying $q=\cdot5$, $n_6'=\cdot010$ instead of $\cdot004$. This is therefore the nearest value. From the expansion of the curve we see that this is correct.

Again consider the curve

$$y = \frac{(\cdot 4)^{\cdot 8} e^{-\cdot 4x} x^{-\cdot 2}}{\Gamma(\cdot 8)},$$

where $h = 2\sqrt{\cdot 8}$. $$\int_0^x y\,dx = \frac{1}{\Gamma(\cdot 8)} \int_0^{\cdot 4x} e^{-v} v^{-\cdot 2} dv = I(u, p),$$

where $$p = -\cdot 2, \quad u = \frac{v}{\sqrt{\cdot 8}} = \frac{\cdot 4x}{\sqrt{\cdot 8}}.$$

We have

x	u	$I(u, p)$	Frequencies
0	0	0	
$2\sqrt{\cdot 8}$	·8	·611,0697	$\cdot 611{,}0697 = n_1'$
$4\sqrt{\cdot 8}$	1·6	·825,3995	$\cdot 214{,}3298 = n_2'$
$6\sqrt{\cdot 8}$	2·4	·919,4155	$\cdot 094{,}0160 = n_3'$
$8\sqrt{\cdot 8}$	3·2	·962,2971	$\cdot 042{,}8816 = n_4'$
$10\sqrt{\cdot 8}$	4·0	·982,2136	$\cdot 019{,}9165 = n_5'$
$12\sqrt{\cdot 8}$	4·8	·991,5622	$\cdot 009{,}3486 = n_6'$

$q = \cdot 2$ gives $n_6' = -\cdot 030$,
$q = \cdot 8$ $\quad + \cdot 021$,
$q = \cdot 6$ $\quad - \cdot 031$,
$q = \cdot 7$ $\quad - \cdot 011$.

The true value of $n_6' = \cdot 009$, therefore $q = \cdot 8$ is the correct value to use, which is as it should be.

If, however, in any case it is found that the selected value for q does not give a good fit, it is as well to try another value. Now it is obvious that $q - 1$ should agree with the m_1 of the Pearson Type curve fitted. If therefore a bad fit is obtained and this is not the case, the value next tried for q might well be the value obtained for m_1 increased by unity. As a general rule in fitting a curve the start should be fixed at zero since our auxiliary curve assumes an asymptote there. Only three moments will therefore be required, and q might be selected to make the fourth moment of the curve, obtained with fixed start, agree with that obtained from the frequencies. The disadvantage of these two methods is that they involve the complete fitting of a curve for several different values of q. They should therefore only be used when the first method has failed.

(4) Let us now illustrate this method by one or two examples.

Example 1. The curve $y = \frac{(\cdot 4)^{\cdot 8} e^{-\cdot 4x} x^{-\cdot 2}}{\Gamma(\cdot 8)}$ between $x = 0$ and $x = 26\sqrt{\cdot 8}$ for thirteen equal subranges.

x	u	$I(u, p)$	Frequencies	Distances of Centres from $x = 2\sqrt{\cdot 8}$
0	0	0	·611,0697	
$2\sqrt{\cdot 8}$	·8	·611,0697	·214,3298	$\sqrt{\cdot 8}$
$4\sqrt{\cdot 8}$	1·6	·825,3995	·094,0160	$3\sqrt{\cdot 8}$
$6\sqrt{\cdot 8}$	2·4	·919,4155	·042,8816	$5\sqrt{\cdot 8}$
$8\sqrt{\cdot 8}$	3·2	·962,2971	·019,9165	$7\sqrt{\cdot 8}$
$10\sqrt{\cdot 8}$	4·0	·982,2136	·009,3486	$9\sqrt{\cdot 8}$
$12\sqrt{\cdot 8}$	4·8	·991,5622	·004,4187	$11\sqrt{\cdot 8}$
$14\sqrt{\cdot 8}$	5·6	·995,9809	·002,0988	$13\sqrt{\cdot 8}$
$16\sqrt{\cdot 8}$	6·4	·998,0797	·001,0006	$15\sqrt{\cdot 8}$
$18\sqrt{\cdot 8}$	7·2	·999,0803	·000,4784	$17\sqrt{\cdot 8}$
$20\sqrt{\cdot 8}$	8·0	·999,5587	·000,2292	$19\sqrt{\cdot 8}$
$22\sqrt{\cdot 8}$	8·8	·999,7879	·000,1100	$21\sqrt{\cdot 8}$
$24\sqrt{\cdot 8}$	9·6	·999,8979	·000,0529	$23\sqrt{\cdot 8}$
$26\sqrt{\cdot 8}$	10·4	·999,9508		

For crude moments about $x = 2\sqrt{\cdot 8}$ we have

$$N_{2p}\nu_1'' = 1{\cdot}041{,}2521\sqrt{\cdot 8}, \qquad N_{2p}\nu_2'' = 4{\cdot}224{,}3174,$$

$$N_{2p}\nu_3'' = \cdot 8\sqrt{\cdot 8}\,(41{\cdot}213{,}6353), \quad N_{2p}\nu_4'' = \cdot 64\,(424{\cdot}802{,}2139),$$

and $\frac{1}{12}N_{2p}h^2 = \cdot 103{,}7016, \qquad \frac{1}{4}N_{2p}h^2\nu_1'' = \cdot 8\sqrt{\cdot 8}\,(1{\cdot}041{,}2521),$

$$\tfrac{1}{2}N_{2p}h^2\nu_2'' = 6{\cdot}758{,}9078, \qquad \tfrac{7}{240}N_{2p}h^4 = \cdot 298{,}6667.$$

We have already found that the best value for q is ·8,

$$K_1 = -\cdot 402{,}8216, \quad K_2 = \cdot 279{,}8478, \quad K_3 = -\cdot 221{,}5659, \quad K_4 = \cdot 196{,}4822.$$

Therefore the corrected moments about $x = 2\sqrt{\cdot 8}$ are

$$N\mu_1' = \cdot 235{,}6090\sqrt{\cdot 8}, \qquad N\mu_2' = 5{\cdot}016{,}129,$$

$$N\mu_3' = \sqrt{\cdot 8}\,(30{\cdot}719{,}8851), \quad N\mu_4' = 267{\cdot}425{,}1539.$$

These are then transferred to the origin, and in the following table the results are

compared with the true values obtained by integration, and the values obtained when q is taken at its mean value of ·5.

	Value $q = \cdot 5$	Value $q = \cdot 8$	True Value
$N\mu_1''$	1·958,8265	1·999,501	1·998,7352
$N\mu_2''$	8·935,2390	8·969,9203	8·967,1761
$N\mu_3''$	68·932,507	62·143,182	62·139,32
$N\mu_4''$	624·156,46	575·406,86	575·651,48

We see that with the correct value for q, all the four moments are within ·06 °/₀ error of the true values, whereas with the mid-value for q, the error is as great as 11 °/₀.

Example 2. $Z = \sin(x^{\frac{1}{3}})$, $y = \frac{1}{3}x^{-\frac{2}{3}}\cos(x^{\frac{1}{3}})$ from $x = 0$ to $x = \frac{\pi^3}{8}$.

x	$x^{\frac{1}{3}}$	$\sin x^{\frac{1}{3}}$	Frequencies	Distances of Centres from $x = \frac{\pi^3}{64}$
0	0	0		
			·707,1068	
$\pi^3/64$	45°	·707,1068		
			·128,6665	$\pi^3/128$
$2\pi^3/64$	56°41′·7867	·835,7733		
			·069,8046	$3\pi^3/128$
$3\pi^3/64$	64°54′·0739	·905,5779		
			·042,3743	$5\pi^3/128$
$4\pi^3/64$	71°25′·9830	·947,9522		
			·026,2169	$7\pi^3/128$
$5\pi^3/64$	76°56′·9349	·974,1691		
			·015,5334	$9\pi^3/128$
$6\pi^3/64$	81°46′·2256	·989,7025		
			·007,9603	$11\pi^3/128$
$7\pi^3/64$	86°4′·9142	·997,6628		
			·002,3372	$13\pi^3/128$
$8\pi^3/64$	90°	1		

Trying $q = \cdot 3$, Table I gives $n_6' = \cdot 014$,

$q = \cdot 4$, $\cdot 029$.

True $n_6' = \cdot 016$. By linear interpolation our best value is $q = \cdot 31$.

From the expansion for the sine we see that q is really $\frac{1}{3}$. The best we can do is to take $q = \cdot 3$.

The crude moments about $x = \frac{\pi^3}{64}$ are

$$N_{2p}\nu_1'' = \cdot 991,2176\,\frac{\pi^3}{128}, \quad N_{2p}\nu_2'' = \frac{\pi^6}{(128)^2}(5\cdot 717,2820),$$

$$N_{2p}\nu_3'' = \frac{\pi^9}{(128)^3}(43\cdot 356,4112),$$

$$K_1 = -\cdot 568,1722, \quad K_2 = \cdot 482,9315^-, \quad K_3 = -\cdot 437,2316.$$

Whence we have the following comparative table for moments about $x = 0$:

	True Value	Value $q = \cdot 3$	Error
μ_1	·451,008	·449,318	·4 °/₀
μ_2	·644,001	·643,755	·04 °/₀
μ_3	1·285,53	1·285,44	·007 °/₀

If we take $q = \cdot 31$ and use linear interpolation for K_1 the error is halved, and if we take $q = \frac{1}{3}$, the error in the mean becomes ·13 °/₀.

These results seem to show that a table for the argument-interval $\frac{1}{10}$ is sufficient for most purposes, but this may not be so between $q = 0$ and ·1, where the differences are much larger. To decide this let us take a curve for which the true value of q is ·05. With our tables as they stand at present, the best we could do would be to take q as ·1.

Example 3. $y = \dfrac{\cdot 2^{\cdot 05} e^{-\cdot 2x} x^{-\cdot 95}}{\Gamma(\cdot 05)}$ from $x = 0$ to $x = \dfrac{\sqrt{20}}{4}$ for ten equal intervals.

$$Z = \int_0^{x_1} \frac{\cdot 2^{\cdot 05} e^{-\cdot 2x} x^{-\cdot 95}}{\Gamma(\cdot 05)} dx = \int_0^{\cdot 2x_1} \frac{e^{-v} v^{-\cdot 95} dv}{\Gamma(\cdot 05)} = I(u, p),$$

where $p = -\cdot 95$, $u\sqrt{\cdot 05} = \cdot 2x_1$.

x	u	$I(u, p)$	Frequencies	Distances of Centres from $x = \sqrt{20}/40$
0	0	0		
			·848,5479	
$\sqrt{20}/40$	·1	·848,5479		
			·029,0035	$\sqrt{20}/40$
$2\sqrt{20}/40$	·2	·877,5514		
			·017,0436	$3\sqrt{20}/40$
$3\sqrt{20}/40$	·3	·894,5950		
			·012,0294	$5\sqrt{20}/40$
$4\sqrt{20}/40$	·4	·906,6244		
			·009,2406	$7\sqrt{20}/40$
$5\sqrt{20}/40$	·5	·915,8650		
			·007,4578	$9\sqrt{20}/40$
$6\sqrt{20}/40$	·6	·923,3228		
			·006,2179	$11\sqrt{20}/40$
$7\sqrt{20}/40$	·7	·929,5407		
			·005,3048	$13\sqrt{20}/40$
$8\sqrt{20}/40$	·8	·934,8455		
			·004,6045	$15\sqrt{20}/40$
$9\sqrt{20}/40$	·9	·939,4500		
			·004,0502	$17\sqrt{20}/40$
$10\sqrt{20}/40$	1·0	·943,5002		

The true value of the first moment about $x = 0$ is ·045,2901.

Testing $q = \cdot 1$, Table I gives $n_6' = \cdot 012$,

$q = 0$, ·004.

By linear interpolation, $q = \cdot043$.

$q = \cdot1$ gives the first moment about $x = 0$ as $\cdot047,0607$, i.e. 4 °/₀ error,

$q = \cdot043$ and linear interpolation the result is $\cdot045,0602$, i.e. $\cdot34$ °/₀ error,

$q = \cdot05$ and linear interpolation the result is $\cdot045,3058$, i.e. $\cdot03$ °/₀ error.

This seems to indicate that an extension to our tables to the argument-interval $\frac{1}{100}$ between $q = 0$ and $q = \cdot1$ would be useful, and it has therefore been computed. Using the extension to Table I, $q = \cdot05$ gives $n_6' = \cdot007,457$, which is very good.

Then the first moment is $\cdot045,3076$, i.e. $\cdot04$ °/₀ error.

Example 4. Let us now apply our corrections to data obtained from runs in coin-tossing*.

Runs	Frequencies	Runs	Frequencies
1	4165	10	9
2	2028	11	2
3	982	12	1
4	480	13	1
5	266	14	1
6	132	15	1
7	71		
8	36		
9	17		8192

Here $n_6 = 132$.

Table I gives

$$q = \cdot7, \quad n_6 = -253\cdot6,$$
$$q = \cdot9, \quad n_6 = +34\cdot8,$$
$$q = 1\cdot0, \quad n_6 = +375,$$

whence $q = \cdot93$.

We therefore take $q = \cdot9$.

Table IV then gives

$$K_1 = -2604\cdot306,99, \quad K_2 = 1682\cdot917,97,$$
$$K_3 = -1267\cdot010,60, \quad K_4 = 1119\cdot669,46.$$

Taking moments about $x = 1$ for all frequencies except the first, we have

$$4027\mu_1'' = 6107\cdot5, \quad 4027\mu_2'' = 17,684\cdot75,$$
$$4027\mu_3'' = 77,705\cdot875, \quad 4027\mu_4'' = 451,458\cdot6875,$$

whence equations (VIII) give

$$8192\mu_1'' = 3503\cdot193,01, \quad \therefore \mu_1'' = \cdot427,636,$$
$$8192\mu_2'' = 19,032\cdot084,64, \quad \mu_2'' = 2\cdot323,253,$$
$$8192\mu_3'' = 74,911\cdot989,40, \quad \mu_3'' = 9\cdot144,530,$$
$$8192\mu_4'' = 443,853\cdot436,13, \quad \mu_4'' = 54\cdot181,328.$$

* The data are the sum of results experimentally determined by Buffon and by three friends of De Morgan. The latter gave them in the *Assurance Magazine*, Vol. x. 1861, p. 251. I have adopted Palin Elderton's smoothing of the last four entries (*Frequency Curves*, 2nd ed. p. 106).

The mean is, therefore, 1·427,636 and the moments about it are

$$\nu_2 = 2{\cdot}140{,}379, \quad \nu_3 = 6{\cdot}320{,}416, \quad \nu_4 = 40{\cdot}988{,}041.$$

Therefore $$\beta_1 = 4{\cdot}073{,}98, \quad \beta_2 = 8{\cdot}946{,}97.$$

This is a Pearson Type I J-curve.

If we fix the start we have $r = \dfrac{2(\mu_1^2\mu_2 + \mu_1\mu_3 - \mu_2^2)}{2\mu_2^2 - \mu_1\mu_3}$, where μ_1 stands for the mean and the other coefficients for moments about it. Whence $r = 126{\cdot}508{,}670$. This would really give $\beta_2 = 9{\cdot}017{,}46$ instead of 8·946,97, a difference of about half of 1 %. With the usual notation we then have

$$e = 117{\cdot}756{,}00, \quad m_2 = 124{\cdot}570{,}905, \quad m_1 = -{\cdot}062{,}235.$$

We see that m_1 agrees very closely with $q - 1 (= -{\cdot}07)$,

$$b = 192{\cdot}594{,}54, \quad a_1 = -{\cdot}096{,}268, \quad a_2 = 192{\cdot}691{,}259.$$

The equation of our curve with origin at the start is

$$y = \frac{N\Gamma(m_1 + m_2 + 2)}{b\,\Gamma(m_1 + 1)\,\Gamma(m_2 + 1)} \left(\frac{x}{b}\right)^{m_1} \left(1 - \frac{x}{b}\right)^{m_2},$$

i.e. $$y = 4052{\cdot}866 \left(\frac{x}{b}\right)^{-{\cdot}062{,}235} \left(1 - \frac{x}{b}\right)^{124{\cdot}570{,}905},$$

where b has the value given above.

This curve has been drawn as Diagram I, and we can see by eye that it gives a pretty good fit.

The frequencies are calculated from the formulae given on page 41 of *Tracts for Computers*, No. 7, and we have the following results:

True Value	Calculated Value	Contribution to χ^2
4165	4183·5	·082
2028	1982·9	1·026
982	997·6	·244
480	506·2	1·356
266	258·5	·218
132	129·7	·041
71	66·2	·348
36	33·5	·187
17, 9, 2, 1, 1, 1, 1 (=32)	16·9, 8·6, 4·2, 2·2, 2 (=33·9)	·106
		3·608

The tables for testing Goodness of Fit give $P = {\cdot}89$, a very good result.

RUNS IN COIN TOSSING

$$y = 4052{\cdot}866\left(\frac{x}{192{\cdot}5945}\right)^{-{\cdot}0622}\left(1-\frac{x}{192{\cdot}5945}\right)^{124{\cdot}5709}$$

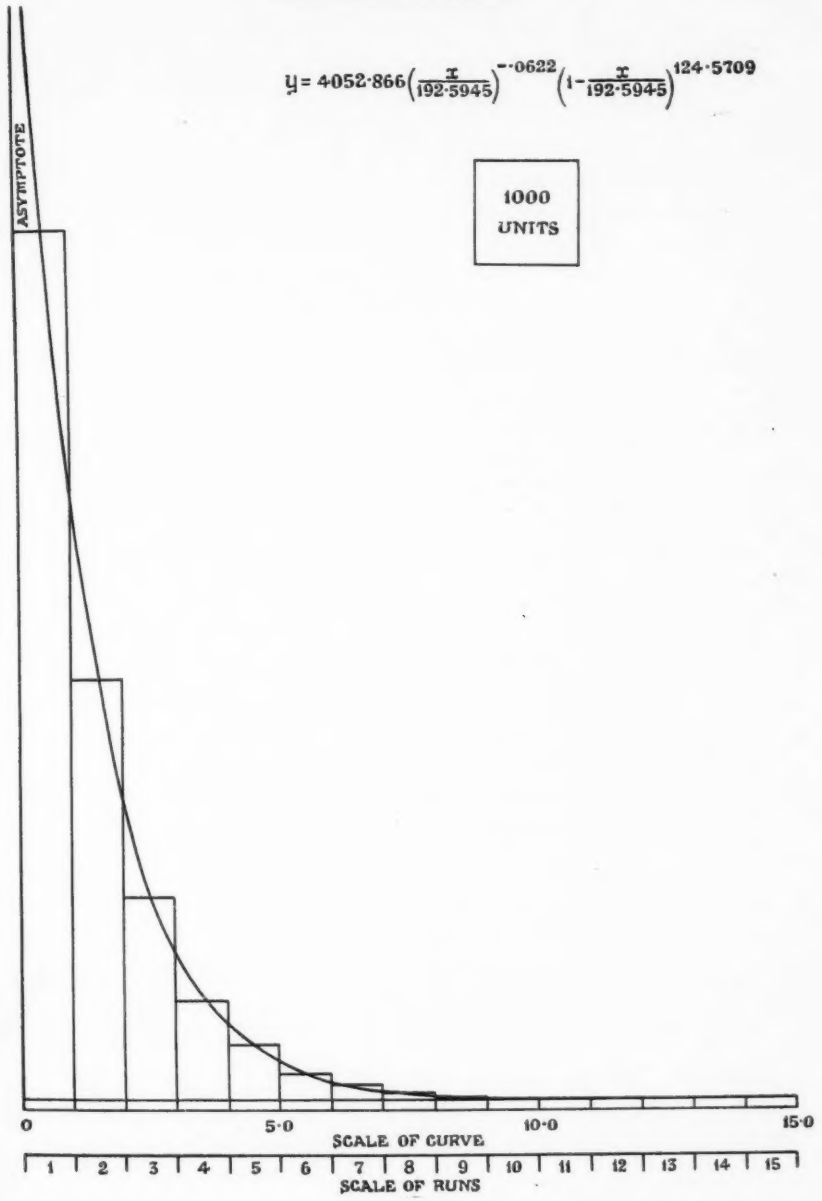


Diagram I.

Example 5. As a further example of a J-curve, we will use the data for De Vries' Buttercup Petals.

No. of Petals	Frequency
5	133
6	55
7	23
8	7
9	2
10	2
Total	222

Here $n_6 = 2$.

Table I gives

$q = \cdot05$, $n_6 = 2\cdot22$,
$q = \cdot06$, $n_6 = 2\cdot06$,
$q = \cdot10$, $n_6 = 1\cdot42$,
$q = \cdot70$, $n_6 = -\ \cdot06$,
$q = \cdot80$, $n_6 = 5\cdot10$,

whence $q = \cdot06$ or $\cdot75$.

Taking $q = \cdot06$ we find $m_1 = -\cdot253$, whereas $q - 1 = -\cdot94$. Therefore let us take $q = \cdot75$, interpolating linearly in Table IV. We might here reasonably expect difficulty in selecting q owing to the large probable errors of the last frequencies.

We have

$K_1 = -86\cdot806{,}795$, $K_2 = 59\cdot403{,}511$, $K_3 = -46\cdot652{,}557$, $K_4 = 41\cdot657{,}089$;
$89\mu_1'' = 95\cdot5$, $89\mu_2'' = 174\cdot25$, $89\mu_3'' = 461\cdot875$, $89\mu_4'' = 1513\cdot5625$;
$222\mu_1'' = 8\cdot693{,}205$, $222\mu_2'' = 226\cdot236{,}844$, $222\mu_3'' = 391\cdot347{,}443$,
$222\mu_4'' = 1470\cdot690{,}422$; $\mu_1'' = \cdot039{,}159$, $\mu_2'' = 1\cdot019{,}085^-$,
$\mu_3'' = 1\cdot762{,}826$, $\mu_4'' = 6\cdot624{,}732$.

Therefore the mean is $1\cdot039{,}159$, and moments about it are

$$\nu_2 = 1\cdot017{,}552, \quad \nu_3 = 1\cdot643{,}229, \quad \nu_4 = 6\cdot357{,}982,$$

whence $\beta_1 = 2\cdot562{,}871$ and $\beta_2 = 6\cdot140{,}533$.

This is again a Pearson Type I J-curve.

Fixing the start we have $\nu_4 = 9\cdot750{,}760$ (whence $\beta_2 = 6\cdot072{,}250^+$ instead of $6\cdot140{,}533$, a difference of about 1 %),

$e = 7\cdot774{,}595^+$, $m_2 = 7\cdot874{,}726$, $m_1 = -\cdot123{,}962$ $(q - 1 = -\cdot25)$,
$b = 11\cdot566{,}384$, $a_1 = -\cdot184{,}987$, $a_2 = 11\cdot751{,}371$.

The equation of the curve with origin at the start is therefore

$$y = 118\cdot6404 \left(\frac{x}{b}\right)^{-\cdot123{,}962} \left(1 - \frac{x}{b}\right)^{7\cdot874{,}726}.$$

DE VRIES' BUTTERCUP PETALS

$$y = 118{\cdot}6404\left(\cdot 0865\bar{x}\right)^{-\cdot 1240}\left(1-\cdot 0865\bar{x}\right)^{7\cdot 8747}$$

$$q = \cdot 75$$

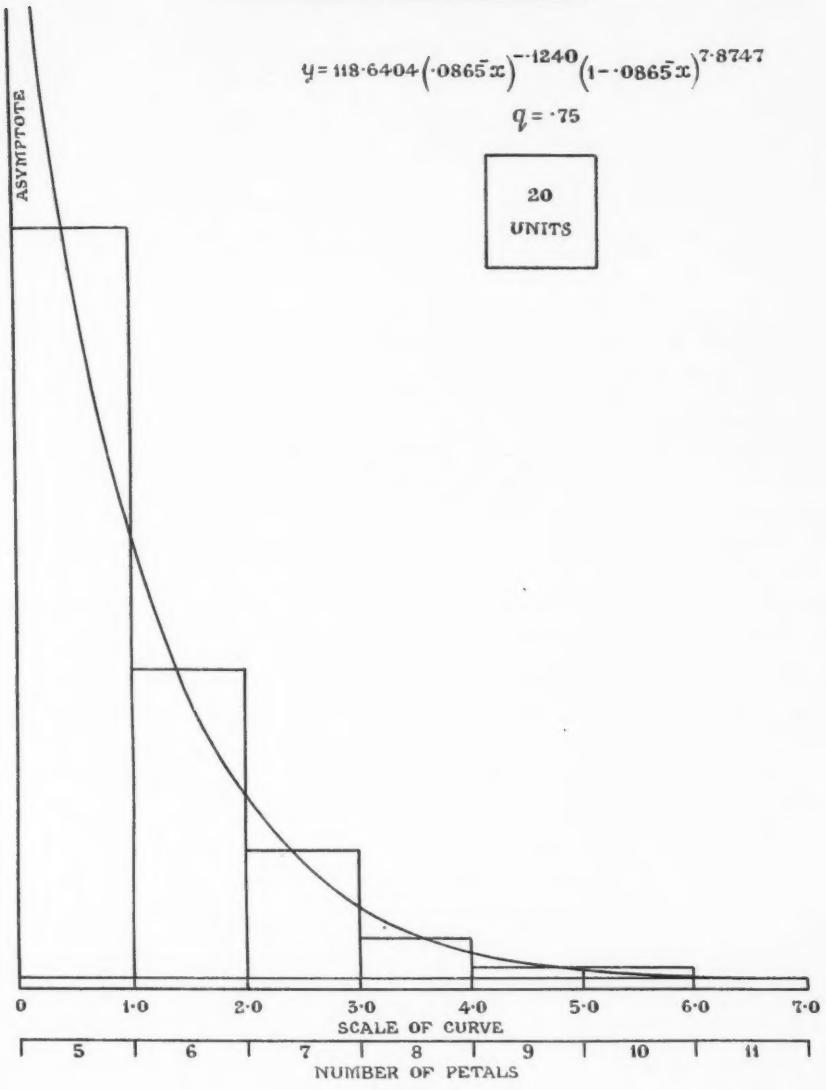


Diagram II.

This has been drawn and we can see by eye that it gives a good fit. The frequency from 0 to x is $222I_{x/b}$ (·876,038, 8·874,726) in the notation of page 5 of *Tracts for Computers*, No. 7. We have therefore the following table:

True Value	Calculated Value	Contribution to χ^2
133	135·1	·033
55	52·8	·092
23	21·7	·078
7	8·4	·233
2, 2, 0 } =4	2·9, ·9, ·2 } =4·0	·000
	Total	·436

Testing for Goodness of Fit this gives $P = ·96$, a very much better fit than the values found by other methods.

We notice that in this case the value indicated for q by m_1 is ·9, and it will be interesting to see what result we obtain if we take this value instead of ·75.

The mean is now 1·049,392 and the moments about it are

$$\nu_2 = 1·001,356, \quad \nu_3 = 1·631,200, \quad \nu_4 = 6·270,489,$$

whence $\beta_1 = 2·650,017$ and $\beta_2 = 6·253,517$.

Fixing the start we have $r = 12·339,241$ (whence $\beta_2 = 4·368,963$, a wide divergence),

$e = 10·713,281$, $m_2 = 10·399,433$, $m_1 = -·060,192$ $(q - 1 = -·1)$, $b = 13·984,386$.

The equation of the curve with origin at the start is

$$y = 150·14 \left(\frac{x}{b}\right)^{-·060,192} \left(1 - \frac{x}{b}\right)^{10·399,433},$$

where $\frac{1}{b} = ·07151$.

This has been drawn, but the difference between it and the former curve is scarcely noticeable by eye.

The frequencies calculated as before are

True Value	Calculated Value	Contribution to χ^2
133	132·6	·001
55	54·7	·002
23	22·3	·022
7	8·5	·265
2, 2, 0 } 4	3·9	·003
	Total	·293

Testing for Goodness of Fit this gives $P = \cdot 97$.

By making the value of q agree more closely with that of $1 + m_1$ we have obtained a better fit, though in this case the original was so good as to be hardly worth improving.

DE VRIES' BUTTERCUP PETALS

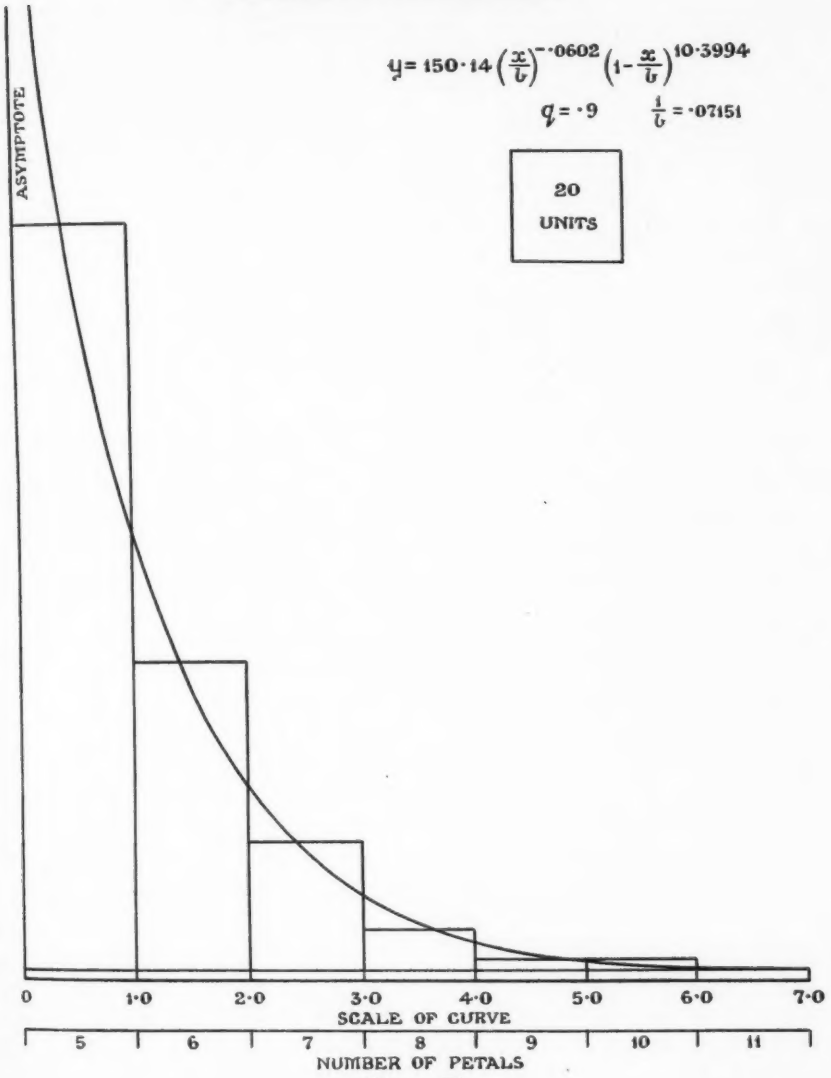

$y = 150 \cdot 14 \left(\frac{x}{b}\right)^{-\cdot 0602} \left(1 - \frac{x}{b}\right)^{10 \cdot 3994}$
$q = \cdot 9$ $\frac{1}{b} = \cdot 07151$
20 UNITS
ASYMPTOTE
0 1·0 2·0 3·0 4·0 5·0 6·0 7·0
SCALE OF CURVE
5 6 7 8 9 10 11
NUMBER OF PETALS

Diagram III.

Example 6. For De Vries' High Blossoms in Clover we have

No. of High Blossoms	Frequency	No. of High Blossoms	Frequency
0	325	6	18
1	83	7	7
2	66	8	6
3	51	9	1
4	36	10	1
5	36		
		Total	630

Here $n_6 = 36$.

Table I gives $q = \cdot3$, $n_6 = 28\cdot182$,

$q = \cdot4$, $n_6 = 35\cdot990$,

$q = \cdot5$, $n_6 = 47\cdot887$.

Therefore we take $q = \cdot4$.

We have

$K_1 = -252\cdot283{,}019$, $K_2 = 207\cdot661{,}343$, $K_3 = -183\cdot734{,}737$, $K_4 = 170\cdot970{,}121$;

$305\mu_1'' = 763\cdot5$, $305\mu_2'' = 2998\cdot25$, $305\mu_3'' = 14773\cdot875$, $305\mu_4'' = 83812\cdot0625$.

Therefore

$$630\mu_1'' = 511\cdot216{,}981, \quad 630\mu_2'' = 3180\cdot494{,}676,$$
$$630\mu_3'' = 14399\cdot265{,}263, \quad 630\mu_4'' = 82602\cdot803{,}454.$$

Therefore $\mu_1'' = \cdot811{,}455{,}53$, $\mu_2'' = 5\cdot048{,}404{,}25^-$,

$\mu_3'' = 22\cdot855{,}976{,}61$, $\mu_4'' = 131\cdot115{,}561{,}04$.

Therefore the mean is $1\cdot811{,}455{,}53$, and the moments about it are

$$\nu_2 = 4\cdot389{,}944{,}17, \quad \nu_3 = 11\cdot634{,}932{,}05, \quad \nu_4 = 75\cdot573{,}4535^+,$$

whence $\beta_1 = 1\cdot600{,}113$, $\beta_2 = 3\cdot921{,}491$.

This is again a Pearson Type I J-curve.

Fixing the start we have $r = 1\cdot856{,}019$ (whence $\beta_2 = 3\cdot670$ instead of $3\cdot921$),

$b = 8\cdot732{,}829$, $e = \cdot566{,}337$, $m_2 = \cdot471{,}024$, $m_1 = -\cdot615{,}005$ $(q - 1 = -\cdot6)$,

$$y_0 = 33\cdot4522.$$

The equation of the curve with origin at the start is therefore

$$y = 33\cdot4522\,(\cdot114{,}510x)^{-\cdot6150+}\,(1 - \cdot114{,}510x)^{\cdot4710},$$

which is drawn in the accompanying Diagram IV.

DE VRIES' HIGH BLOSSOMS IN CLOVER

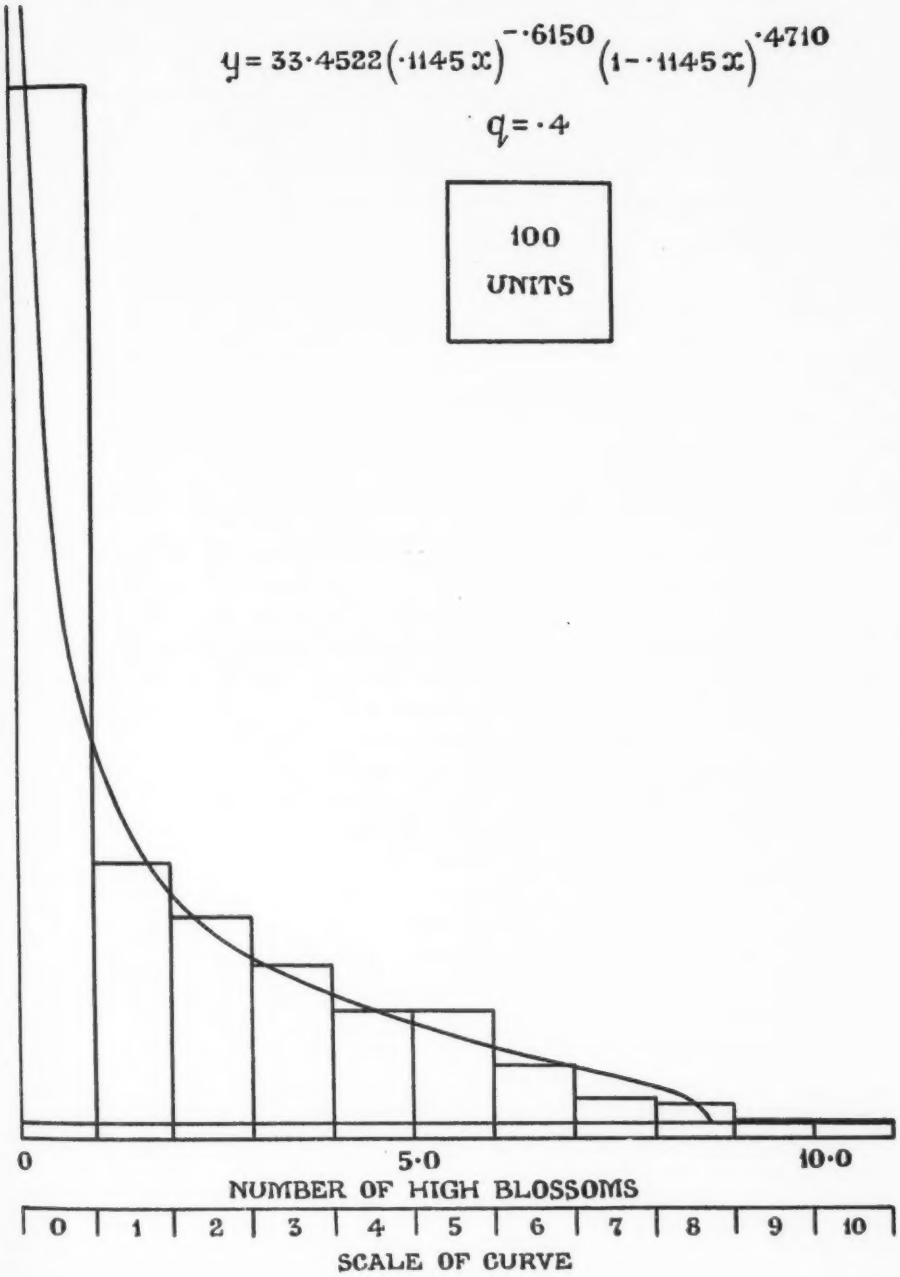


Diagram IV.

The frequencies calculated as before are

True Value	Calculated Value	Contribution to χ^2
325	324	·003
83	92	·880
66	61	·410
51	48	·188
36	36	0
36	28	2·286
18	21	·429
7, 6, 1, 1 } = 15	20	1·250
	Total	$\chi^2 = 5{\cdot}446$ $n' = 8$

Testing for Goodness of Fit this gives $P = {\cdot}61$.

(5) *Curves asymptotic at both terminals.*

Let us now extend our work to the case of curves which are asymptotic at both terminals, that is, to U-shaped curves.

Let n_p', n'_{p-1}, n'_{p-2}, n'_{p-3}, n'_{p-4} be the relative frequencies of the last five subranges, the total number of subranges being p.

We take for our auxiliary curve at the second terminal

$$Z = N(x_p - x)^q [A' + B'(x_p - x) + C'(x_p - x)^2 + D'(x_p - x)^3 + E'(x_p - x)^4],$$

where $x_p = ph$.

Therefore
$$n_p' = A' + B' + C' + D' + E',$$
$$n_p' + n'_{p-1} = 2^q (A' + 2B' + 2^2 C' + 2^3 D' + 2^4 E'),$$

with three other equations similar to Equations (I).

Accordingly the equations for A', B', C', D' and E' will be the same as equations (II), substituting $-n_p'$ for n_1', $-n'_{p-1}$ for n_2', etc.

The moments of the last frequency about $x = x_{p-1}$ are given by

$$n_p \mu_s'' = \frac{1}{N} \int_{x_{p-1}}^{x_p} (x_{p-1} - x)^s \frac{dZ}{dx}\, dx,$$

whence, if we substitute n_p for n_1, n_{p-1} for n_2, etc. in Table I, we obtain the values of $n_p \mu_s''$ if we change the sign for the first and third moments and leave it unaltered for the second and fourth.

For the abruptness coefficients we have

$$b_s = \frac{1}{N} (-1)^s \left[\frac{d^s Z}{dx^s}\right]_{x = x_{p-1}},$$

and these can therefore be found from Table III by writing n_p for n_1, n_{p-1} for n_2, etc. without change of sign.

We have now to find the corrections necessary when the moments are calculated about $x = 1$:

$$n_p\mu_s'' \text{ about } x=1 = n_p\mu_s'' + s(x_p-2)\,n_p\mu''_{s-1} + \frac{s(s-1)}{2}(x_p-2)^2\,n_p\mu''_{s-2} + \text{etc.},$$

the moments on the right-hand side being calculated about $x = x_{p-1}$.

Let T_1, T_2, T_3, T_4 be the values found for K_1, K_2, K_3, K_4 of Table IV, when n_p is written for n_1, n_{p-1} for n_2, etc. and all signs are changed.

Then moments about $x = 1$ are given by

$$\begin{aligned}
N\mu_1'' &= N_{2,p-1}\nu_1'' + \frac{h}{12}\left(a_1 - \tfrac{1}{60}a_3 + \tfrac{1}{2520}a_5 + b_1 - \tfrac{1}{60}b_3 + \tfrac{1}{2520}b_5\right) \\
&\quad + hn_1\mu_1'' + hn_p\mu_1'' + h(p-2)\,n_p \\
&= N_{2,p-1}\nu_1'' + hK_1 + hT_1 + h(p-2)\,n_p, \\
N\mu_2'' &= N_{2,p-1}(\nu_2'' - \tfrac{1}{12}h^2) + h^2\{-\tfrac{1}{120}(a_2 - \tfrac{5}{126}a_4) + \tfrac{1}{120}(b_2 - \tfrac{5}{126}b_4) \\
&\quad + \tfrac{1}{6}(p-2)(b_1 - \tfrac{1}{60}b_3 + \tfrac{1}{2520}b_5) + n_1\mu_2'' + n_p\mu_2'' \\
&\quad + 2(p-2)\,n_p\mu_1'' + (p-2)^2 n_p\} \\
&= N_{2,p-1}(\nu_2'' - \tfrac{1}{12}h^2) + h^2K_2 + h^2T_2 + 2(p-2)\,h^2T_1 + h^2(p-2)^2 n_p, \\
N\mu_3'' &= N_{2,p-1}(\nu_3'' - \tfrac{1}{4}h^2\nu_1'') + h^3\{-\tfrac{1}{40}(a_1 - \tfrac{5}{63}a_3 + \tfrac{1}{2340}a_5) \\
&\quad - \tfrac{1}{40}(b_1 - \tfrac{5}{63}b_3 + \tfrac{1}{2340}b_5) + \tfrac{1}{40}(p-2)(b_2 - \tfrac{5}{126}b_4) \\
&\quad + \tfrac{1}{4}(p-2)^2(b_1 - \tfrac{1}{60}b_3 + \tfrac{1}{2520}b_5) + n_1\mu_3'' + n_p\mu_3'' + 3(p-2)\,n_p\mu_2'' \\
&\quad + 3(p-2)^2 n_p\mu_1'' + (p-2)^3 n_p\} \\
&= N_{2,p-1}(\nu_3'' - \tfrac{1}{4}h^2\nu_1'') + h^3K_3 + h^3T_3 + 3(p-2)\,T_2h^3 \\
&\quad + 3(p-2)^2\,T_1h^3 + n_p(p-2)^3\,h^3.
\end{aligned}$$

Similarly,

$$N\mu_4'' = N_{2,p-1}(\nu_4'' - \tfrac{1}{2}h^2\nu_2'' + \tfrac{7}{2340}h^4) + h^4[K_4 + T_4 + 4(p-2)\,T_3 + 6(p-2)^2T_2 + 4(p-2)^3\,T_1 + n_p(p-2)^4].$$

We therefore have the following formula for U-curves.

In values tabled for K_1 to K_4 (Table IV) write n_p for n_1, n_{p-1} for n_2, etc. and *change all signs*. Call the values thus obtained T_1, T_2, T_3, T_4. Then

$$\left.\begin{aligned}
N\mu_1'' &= N_{2,p-1}\nu_1'' + h[K_1 + T_1 + (p-2)\,n_p] \\
N\mu_2'' &= N_{2,p-1}(\nu_2'' - \tfrac{1}{12}h^2) + h^2[K_2 + T_2 + 2(p-2)\,T_1 + (p-2)^2 n_p] \\
N\mu_3'' &= N_{2,p-1}(\nu_3'' - \tfrac{1}{4}h^2\nu_1'') + h^3[K_3 + T_3 + 3(p-2)\,T_2 \\
&\qquad + 3(p-2)^2\,T_1 + (p-2)^3 n_p] \\
N\mu_4'' &= N_{2,p-1}(\nu_4'' - \tfrac{1}{2}h^2\nu_2'' + \tfrac{7}{2340}h^4) + h^4[K_4 + T_4 + 4(p-2)\,T_3 \\
&\qquad + 6(p-2)^2T_2 + 4(p-2)^3\,T_1 + n_p(p-2)^4]
\end{aligned}\right\} \ldots\text{(IX)},$$

where ν_s's are calculated in h units for the frequency with the two terminal ranges cut off.

(6) We can illustrate this method on data of the frequencies of different degrees of cloudiness as observed at Greenwich during the years 1890—1904 (excluding 1901) for the month of July.

Degrees of Cloudiness	Frequency	Degrees of Cloudiness	Frequency
10	676	4	45
9	148	3	68
8	90	2	74
7	65	1	129
6	55	0	320
5	45		
		Total	1715

For the first terminal,

$$n_1 = 676, \quad n_2 = 148, \quad n_3 = 90, \quad n_4 = 65, \quad n_5 = 55, \quad n_6 = 45.$$

Table I gives $q = \cdot1$, $n_6 = 42\cdot25$; $q = \cdot2$, $n_6 = 45\cdot05$. We therefore take $q = \cdot2$.

For the second terminal,

$$n_{11} = 320, \quad n_{10} = 129, \quad n_9 = 74, \quad n_8 = 68, \quad n_7 = 45, \quad n_6 = 45.$$

Table I gives

$$q = 0, \ n_6 = -61; \quad q = \cdot4, \ n_6 = -94; \quad q = \cdot6, \ n_6 = -106;$$
$$q = \cdot8, \ n_6 = -90; \quad q = 1\cdot0, \ n_6 = -40.$$

We deduce that the curve is abrupt but not asymptotic, and therefore use the method of Part I, page 240, of *Biometrika*, XII.

For moments about $x = 1$, since $h = 1$, we shall have

$$\mu_1'' = \tfrac{1}{1715}(1039 \times \text{crude moment} + K_1) - \frac{k_1}{1039},$$

$$\mu_2'' = \tfrac{1}{1715}(1039 \times \text{crude moment} - \tfrac{1039}{12} + K_2) + \tfrac{1}{1039}(-20k_1 + k_2),$$

$$\mu_3'' = \tfrac{1}{1715}(1039 \times \text{crude moment} - \tfrac{1}{4}\cdot1039\nu_1'' + K_3) + \tfrac{1}{1039}(-k_3 + 30k_2 - 300k_1),$$

$$\mu_4'' = \tfrac{1}{1715}(1039 \times \text{crude moment} - \tfrac{1}{2}\cdot1039\nu_2'' + \tfrac{7}{240}\cdot1039 + K_4)$$
$$+ \tfrac{1}{1039}(k_4 - 40k_3 + 600k_2 - 4000k_1),$$

where K_1, K_2, K_3, K_4 are those given in Table IV and k_1, k_2, k_3, k_4 are those given by Mr Sandon in *Biometrika*, Vol. XVI. pp. 194—5.

$1039\nu_1'' = 6147\cdot5$, $1039\nu_2'' = 48827\cdot75$, $1039\nu_3'' = 418758\cdot875$,

$1039\nu_4'' = 3706220\cdot9375$; $K_1 = -549\cdot271,353$, $K_2 = 471\cdot927,901$,

$K_3 = -432\cdot431,521$, $K_4 = 413\cdot371,501$; $k_1 = -40\cdot144,792$, $k_2 = -3\cdot492,758$;

$k_3 = 11\cdot570,084$, $k_4 = 3\cdot277,480$.

Therefore

$$\mu_1'' = 3\cdot302,911, \quad \mu_2'' = 29\cdot465,079, \quad \mu_3'' = 254\cdot505,376, \quad \mu_4'' = 2299\cdot177,361.$$

Transferring to the mean we have

$$\nu_2 = 18{\cdot}555{,}856, \quad \nu_3 = 34{\cdot}608{,}148, \quad \nu_4 = 508{\cdot}355{,}915.$$

Therefore $\beta_1 = {\cdot}187{,}462$ and $\beta_2 = 1{\cdot}476{,}406.$

This is a Pearson Type I U-curve.

Fitting the curve in the usual manner we have

$$r = \frac{6(\beta_2 - \beta_1 - 1)}{3(\beta_1 - 2\beta_2 + 6)} = {\cdot}480{,}297,$$

$$e = \frac{r^2}{4 + \frac{1}{4}\beta_1 \dfrac{(r+2)^2}{r+1}} = {\cdot}054{,}994,$$

$$b = \sqrt{\frac{\mu_2 r^2 (r+1)}{e}} = 10{\cdot}734{,}177.$$

Distance from mean to mode $= \dfrac{r+2}{r-2}\sqrt{\tfrac{1}{4}\beta_1 \mu_2} = 1{\cdot}521{,}984.$

Therefore mode is at $x = 5{\cdot}824{,}903.$

$$m'^2 - rm' + e = 0.$$

Therefore $m_2 = -{\cdot}708{,}105^+$ and $m_1 = -{\cdot}811{,}598.$

$$a_1 = \frac{bm_1}{m_1 + m_2} = 5\,732{,}590\,; \quad a_2 = \frac{a_1 m_2}{m_1} = 5{\cdot}001{,}586.$$

Therefore curve starts at ${\cdot}092{,}312$ and ends at $10{\cdot}826{,}489.$

$$y_0 = \frac{N m_1^{m_1} m_2^{m_2}}{b(m_1 + m_2)^{m_1 + m_2}} \frac{\Gamma(m_1 + m_2 + 2)}{\Gamma(m_1 + 1)\,\Gamma(m_2 + 1)} = 55{\cdot}9122.$$

Therefore equation of curve is

$$y = 55{\cdot}9122 \left(1 + \frac{x}{5\,732{,}590}\right)^{-{\cdot}811{,}598} \left(1 - \frac{x}{5{\cdot}001{,}586}\right)^{-{\cdot}708{,}105+}$$

with origin at $x = 5{\cdot}824{,}903.$

We have the following table of values:

$x =$	$-\ 5{\cdot}733$,	$-5{\cdot}325^-$,	$-4{\cdot}825^-$,	$-\ 3{\cdot}825^-$,	$-\ 1{\cdot}825^-$,	0 ,
$y =$	∞ ,	$285{\cdot}954$,	$154{\cdot}682$,	$91{\cdot}337$,	$61{\cdot}224$,	$55{\cdot}912$,
$x =$	$+\ 3{\cdot}175^+$,	$+4{\cdot}175^+$,	$+4{\cdot}675^+$,	$+\ 5{\cdot}002$.		
$y =$	$79{\cdot}789$,	$128{\cdot}322$,	$237{\cdot}998$,	∞ .		

The curve is shown in the accompanying Diagram V (p. 338) and evidently gives a very bad fit, but it brings out a point of great interest. We see that if the left portion of the curve could be shifted a half to the right and the right portion a half to the left, we should have quite a good fit. This would reduce the range from 11 to 10, which is the true range of actual observation. It suggests that the first and last frequencies should be taken from 0 to $\frac{1}{2}$ and $9\frac{1}{2}$ to 10 respectively, i.e. half the range of the other frequencies, which would then

CLOUDINESS AT GREENWICH 1890–1904 (LESS 1901)

$$y = 55{\cdot}9122\left(1+\frac{x}{5{\cdot}7326}\right)^{-{\cdot}8116}\left(1-\frac{x}{5{\cdot}0016}\right)^{-{\cdot}7081}$$

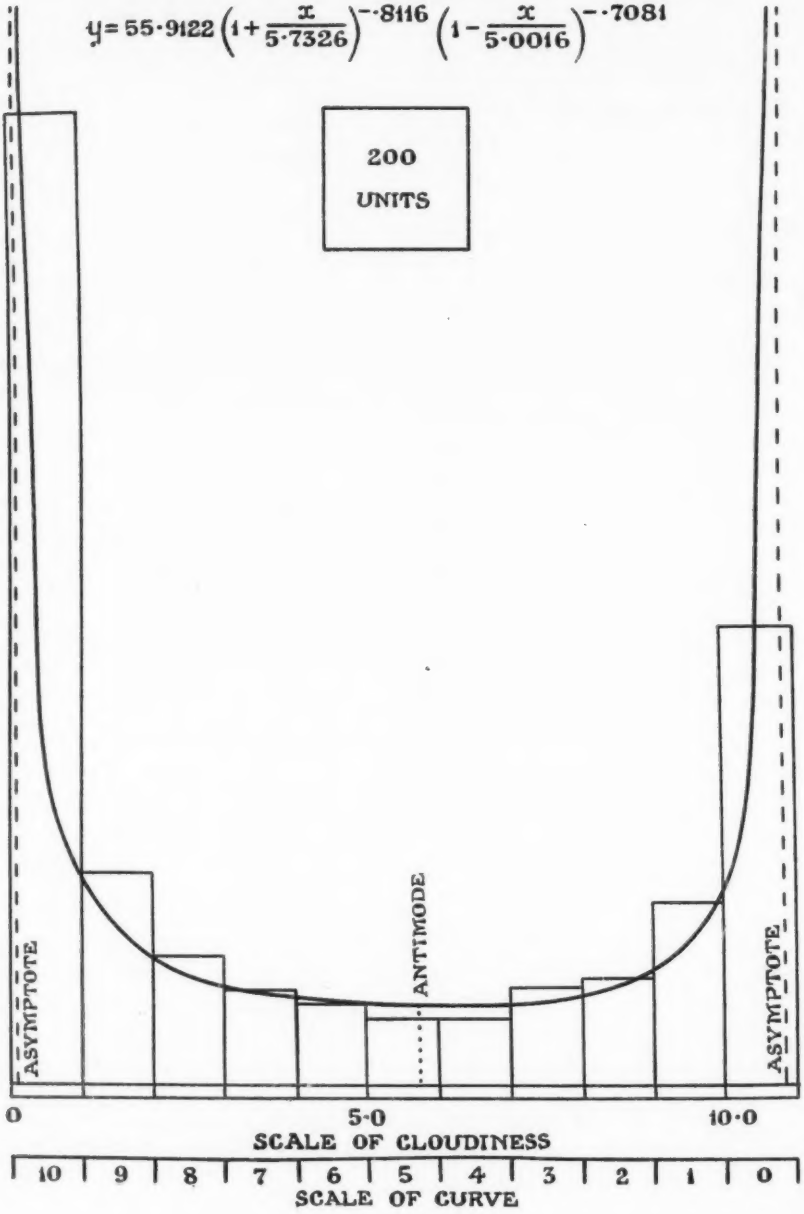


Diagram V.

be taken from $\frac{1}{2}$ to $1\frac{1}{2}$, $1\frac{1}{2}$ to $2\frac{1}{2}$, etc. This seems all the more reasonable since no degrees and 10 degrees of cloudiness presumably represent entirely clear and entirely overcast skies, and therefore should have much smaller ranges than the other degrees. The difficulty is how to apply our method when the two terminal frequencies are given for half the range of the remaining frequencies.

Let us fit the curve

$$Z = N[1 + x^q(A + Bx + Cx^2 + Dx^3 + Ex^4)]$$

to the first six frequencies on ranges 0 to $\frac{1}{2}$, $\frac{1}{2}$ to $1\frac{1}{2}$, etc., thus obtaining A, B, C, D, E and q, and then find the frequencies on ranges $\frac{1}{2}$ to 1, 1 to $1\frac{1}{2}$, etc. by substitution. Treating the other terminal similarly we shall have the whole frequency divided into ranges of a half.

			1st Terminal	2nd Terminal
Let	Nn_1' be frequency from $x = 0$ to $\frac{1}{2}$	=	676	320
	Nn_2' " " $\frac{1}{2}$ to $1\frac{1}{2}$	=	148	129
	Nn_3' " " $1\frac{1}{2}$ to $2\frac{1}{2}$	=	90	74
	Nn_4' " " $2\frac{1}{2}$ to $3\frac{1}{2}$	=	65	68
	Nn_5' " " $3\frac{1}{2}$ to $4\frac{1}{2}$	=	55	45
	Nn_6' " " $4\frac{1}{2}$ to $5\frac{1}{2}$	=	45	45

Then
$$-n_1' = (\cdot5)^q[A + (\cdot5)B + (\cdot5)^2C + (\cdot5)^3D + (\cdot5)^4E],$$
$$-n_1' - n_2' = (1\cdot5)^q[A + (1\cdot5)B + (1\cdot5)^2C + (1\cdot5)^3D + (1\cdot5)^4E].$$

With three similar equations,

.....

.....

$$-n_1' - n_2' - n_3' - n_4' - n_5' - n_6' = (5\cdot5)^q[A + (5\cdot5)B + (5\cdot5)^2C + (5\cdot5)^3D + (5\cdot5)^4E].$$

On solution these give

$$384A = \frac{-105(n_1' + n_2' + \ldots + n_5')}{(4\cdot5)^q} + \frac{540(n_1' + n_2' + \ldots + n_4')}{(3\cdot5)^q} - \frac{1134(n_1' + n_2' + n_3')}{(2\cdot5)^q} + \frac{1260(n_1' + n_2')}{(1\cdot5)^q} - \frac{945n_1'}{(\cdot5)^q},$$

$$24B = \frac{22(n_1' + n_2' + \ldots + n_5')}{(4\cdot5)^q} - \frac{111(n_1' + n_2' + \ldots + n_4')}{(3\cdot5)^q} + \frac{225(n_1' + n_2' + n_3')}{(2\cdot5)^q} - \frac{229(n_1' + n_2')}{(1\cdot5)^q} + \frac{93n_1'}{(\cdot5)^q},$$

$$48C = \frac{-43(n_1' + n_2' + \ldots + n_5')}{(4\cdot5)^q} + \frac{208(n_1' + n_2' + \ldots + n_4')}{(3\cdot5)^q} - \frac{390(n_1' + n_2' + n_3')}{(2\cdot5)^q} + \frac{328(n_1' + n_2')}{(1\cdot5)^q} - \frac{103n_1'}{(\cdot5)^q},$$

$$6D = \frac{2(n_1' + n_2' + \ldots + n_5')}{(4\cdot5)^q} - \frac{9(n_1' + n_2' + \ldots + n_4')}{(3\cdot5)^q} + \frac{15(n_1' + n_2' + n_3')}{(2\cdot5)^q} - \frac{11(n_1' + n_2')}{(1\cdot5)^q} + \frac{3n_1'}{(\cdot5)^q},$$

$$24E = \frac{-(n_1' + n_2' + \ldots + n_5')}{(4{\cdot}5)^q} + \frac{4(n_1' + n_2' + \ldots + n_4')}{(3{\cdot}5)^q} - \frac{6(n_1' + n_2' + n_3')}{(2{\cdot}5)^q} + \frac{4(n_1' + n_2')}{(1{\cdot}5)^q} - \frac{n_1'}{(\cdot 5)^q},$$

and

$$\frac{-192(n_1' + n_2' + \ldots + n_6')}{11^q} = \frac{-3622(n_1' + n_2' + \ldots + n_5')}{9^q} + \frac{13899(n_1' + n_2' + \ldots + n_4')}{7^q} - \frac{21885(n_1' + n_2' + n_3')}{5^q} + \frac{15601(n_1' + n_2')}{3^q} - 4185n_1'.$$

Substituting the values for the first terminal we have

$$n_6 = 11^q\left\{\frac{-1034}{11^q} + \frac{19505{\cdot}979}{9^q} - \frac{70870{\cdot}422}{7^q} + \frac{104181{\cdot}719}{5^q} - \frac{66954{\cdot}292}{3^q} + 14734{\cdot}687\right\}.$$

We find the following values:

$q =$ 0, ·08, ·10, ·12, 1·00,

$n_6 = -436{\cdot}33, \quad -165{\cdot}10, \quad -67{\cdot}75_-, \quad +43{\cdot}52, \quad +57221{\cdot}11.$

Therefore $q = {\cdot}12$ is the best value for q.

Substituting this value for q we have

$$A = -702{\cdot}128{,}820, \quad B = -70{\cdot}292{,}106, \quad C = 10{\cdot}746{,}671,$$
$$D = -{\cdot}315{,}428, \quad E = -{\cdot}082{,}135_-.$$

Frequency up to $x = 1$ is

$$-(A + B + C + D + E) = 762{\cdot}071{,}818.$$

Therefore frequency $\frac{1}{2}$ to 1 is 86·071,818 and 1 to $1\frac{1}{2}$ is 61·928,182.

Frequency up to $x = 2$ is

$$-2^{\cdot 12}(A + 2B + 4C + 8D + 16E) = 873{\cdot}260{,}933.$$

Proceeding in this manner, and treating the second terminal similarly, we obtain the following table of results for range a half:

x	Frequency	x	Frequency
0—1	676·0	10—11	22·5
1—2	86·1	11—12	22·5
2—3	61·9	12—13	22·5
3—4	49·3	13—14	33·1
4—5	40 7	14—15	34·9
5—6	34·6	15—16	35·5
6—7	30·4	16—17	38·5
7—8	27·9	17—18	48·9
8—9	27·1	18—19	80·1
9—10	22·5	19—20	320·0

The value taken for q at the second terminal was ·44.

For the first terminal Table I gives $q = {\cdot}1$, $n_6 = 32{\cdot}79$ and $q = {\cdot}2$, $n_6 = 98{\cdot}05$, so that we take $q = {\cdot}1$.

For the second terminal Table I gives $q = \cdot 4$, $n_6 = 31 \cdot 83$, and $q = \cdot 5$, $n_6 = 62 \cdot 96$, and we take $q = \cdot 4$.

These values agree with those used in splitting up the frequencies, as we should expect. We now have

$$K_1 = -604 \cdot 161,373, \quad T_1 = +242 \cdot 593,046, \quad 719\nu_1'' = 6214 \cdot 1,$$
$$K_2 = +559 \cdot 648,063, \quad T_2 = -195 \cdot 592,259, \quad 719\nu_2'' = 80,492 \cdot 95,$$
$$K_3 = -535 \cdot 911,377, \quad T_3 = +171 \cdot 309,304, \quad 719\nu_3'' = 1,174,477 \cdot 325,$$
$$K_4 = +523 \cdot 571,319, \quad T_4 = -159 \cdot 045,916, \quad 719\nu_4'' = 18,049,426 \cdot 3375.$$
$$N_{2 \ldots p-1} = 719, \quad p - 2 = 18, \quad N = 1715.$$

Therefore moments about $x = 1$ are

$$\mu_1'' = 6 \cdot 771,155^-, \quad \mu_2'' = 112 \cdot 659,148, \quad \mu_3'' = 1903 \cdot 228,954, \quad \mu_4'' = 33,173 \cdot 876,879.$$

On transferring to the mean ($x = 7 \cdot 771,155^-$) we have

$$\nu_2 = 66 \cdot 810,601, \quad \nu_3 = 235 \cdot 626,406, \quad \nu_4 = 6310 \cdot 921,674.$$

Therefore
$$\beta_1 = \cdot 186,171, \quad \beta_2 = 1 \cdot 413,846.$$

This is still a Pearson Type I U-curve.

Fitting as before we have

$$r = \cdot 366,152, \quad e = \cdot 031,960, \quad b = 19 \cdot 567,250^-, \quad m_2 = -\cdot 777,468,$$
$$m_1 = -\cdot 856,379, \quad a_2 = 9 \cdot 311,097, \quad a_1 = 10 \cdot 256,153.$$

Distance, mean to mode $= 2 \cdot 553,758$.

Mode is $x = 10 \cdot 324,914$.

Curve starts at $x = \cdot 068,760$ and ends at $x = 19 \cdot 636,010$.

$y_0 = 24 \cdot 612,09$.

Equation of curve is

$$y = 24 \cdot 612,09 \left(1 + \frac{x}{10 \cdot 256,153}\right)^{-\cdot 856,379} \left(1 - \frac{x}{9 \cdot 311,097}\right)^{-\cdot 777,468}$$

with origin at the mode.

We have the following table of values:

$x = -10 \cdot 256$,	$-9 \cdot 825^-$,	$-9 \cdot 325^-$,	$-8 \cdot 325^-$,	$-6 \cdot 325^-$,	$-3 \cdot 325^-$,	0 ,
$y = \infty$,	$212 \cdot 093$,	$111 \cdot 979$,	$62 \cdot 586$,	$37 \cdot 391$,	$27 \cdot 151$,	$24 \cdot 612$,
$x = +2 \cdot 675^+$,	$+5 \cdot 675^+$,	$+7 \cdot 675^+$,	$+8 \cdot 675^+$,	$+9 \cdot 175^+$,	$+9 \cdot 311$.	
$y = 26 \cdot 261$,	$35 \cdot 553$,	$58 \cdot 956$,	$117 \cdot 316$,	$380 \cdot 607$,	∞ .	

This curve is drawn in the accompanying Diagram VI (p. 342), and we see that we have a much better fit.

To calculate the areas at the two terminals, we use the following formula:

Area up to $x = 1$,

$$A_1 = y_0 b \times \frac{(m_1 + m_2)^{m_1 + m_2}}{m_1^{m_1} m_2^{m_2}} \left(\frac{x}{b}\right)^{1+m_1} \left\{\frac{1}{1+m_1} - \frac{m_2}{2+m_1}\left(\frac{x}{b}\right) - \frac{m_2(1-m_2)}{2(3+m_1)}\left(\frac{x}{b}\right)^2 + \text{etc.}\right\},$$

where x is the range of this first area, namely $\cdot 931,240$.

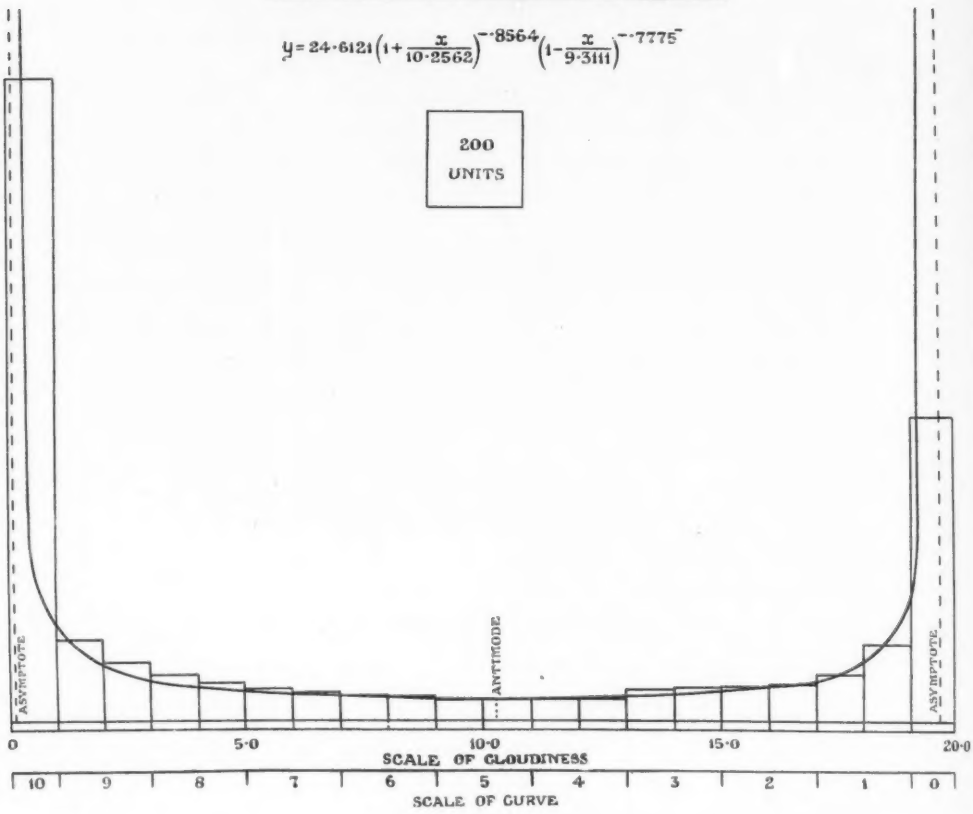


Diagram VI.

Area A_2 from $x = 19$ to the end is the same if we reverse m_1 and m_2, x being now the range of this last area, namely ·636,010. These give

$$A_1 = 705{\cdot}04 \text{ instead of } 676,$$

$$A_2 = 328{\cdot}89 \text{ instead of } 320.$$

Let us see whether we can obtain a still better fit by taking the range as fixed (= 20) and only using the first three moments.

We have $$b^2 = 400 = \frac{\mu_2 r^2 (r+1)}{e} = \mu_2 (r+1)\left[4 + \frac{\frac{1}{4}\beta_1 (r+2)^2}{r+1}\right]$$

$$= \mu_2 [4(r+1) + \tfrac{1}{4}\beta_1 (r+2)^2].$$

Therefore $$\tfrac{1}{4}\beta_1 r^2 + r(\beta_1 + 4) - \frac{400}{\mu^2} = 0.$$

Therefore

$$r = \frac{-(\beta_1+4) \pm \sqrt{4\beta_1+16+\frac{400\beta_1}{\mu_2}}}{\frac{1}{2}\beta_1} = \cdot 428{,}165^-,$$

$$e = \frac{\mu_2 r^2 (r+1)}{400} = \cdot 043{,}731.$$

$m_2 = -\cdot 740{,}086$, $m_1 = -\cdot 831{,}750^-$, $a_2 = 9\cdot 416{,}833$, $a_1 = 10\cdot 583{,}167$.

Distance, mean to mode = 2·724,079.

Mode is 10·495,234.

Curve starts at −·087,933 and ends at 19·912,067.

$y_0 = 27\cdot 437{,}08$.

Equation of curve is

$$y = 27\cdot 437{,}08 \left(1+\frac{x}{10\cdot 583{,}167}\right)^{-\cdot 831{,}750-} \left(1-\frac{x}{9\cdot 416{,}833}\right)^{-\cdot 740{,}086}.$$

We have the following table of values:

$x =$	$-10\cdot 583$,	$-10\cdot 495^+$,	$-9\cdot 995^+$,	$-9\cdot 495^+$,	$-8\cdot 495^+$,
$y =$	∞ ,	847·384 ,	177·792 ,	108·641 ,	65·777 ,
$x =$	$-7\cdot 495^+$,	$-6\cdot 495^+$,	$-5\cdot 495^+$,	$-3\cdot 495^+$,	$-1\cdot 495^+$,
$y =$	49·554 ,	37·608 ,	35·905$^-$,	30·317 ,	27·925$^-$,
$x =$	0 ,	$+2\cdot 505^-$,	4·505$^-$,	6·505$^-$,	7·505$^-$,
$y =$	27·437 ,	28·906 ,	33·069 ,	43·903 ,	57·170 ,
$x =$	8·505$^-$,	9·005$^-$,	9·305$^-$,	9·417 .	
$y =$	94·547 ,	166·619 ,	431·221 ,	∞ .	

The curve is plotted in the accompanying Diagram VII (p. 344).

For the first and last frequencies we have $A_1 = 678\cdot 0$ instead of 676 and $A_2 = 321\cdot 7$ instead of 320.

$$\frac{y_0 b \times (m_1+m_2)^{m_1+m_2}}{m_1^{m_1} m_2^{m_2}} = 185\cdot 080.$$

If x be now measured from the start of the curve, area from $x=0$ to $x=x$ is

$$185\cdot 080 \left(\frac{x}{b}\right)^{1+m_1} \left[\frac{1}{1+m_1} - \frac{m_2}{m_1+2}\frac{x}{b} + \frac{m_2(m_2-1)}{1\,.\,2\,(m_1+3)}\left(\frac{x}{b}\right)^2 - \frac{m_2(m_2-1)(m_2-2)}{1\,.\,2\,.\,3\,(m_1+4)}\left(\frac{x}{b}\right)^3 + \text{etc.}\right],$$

with a similar expression when x is measured from the end of the curve. The areas calculated from these formulae are given in the table on p. 345.

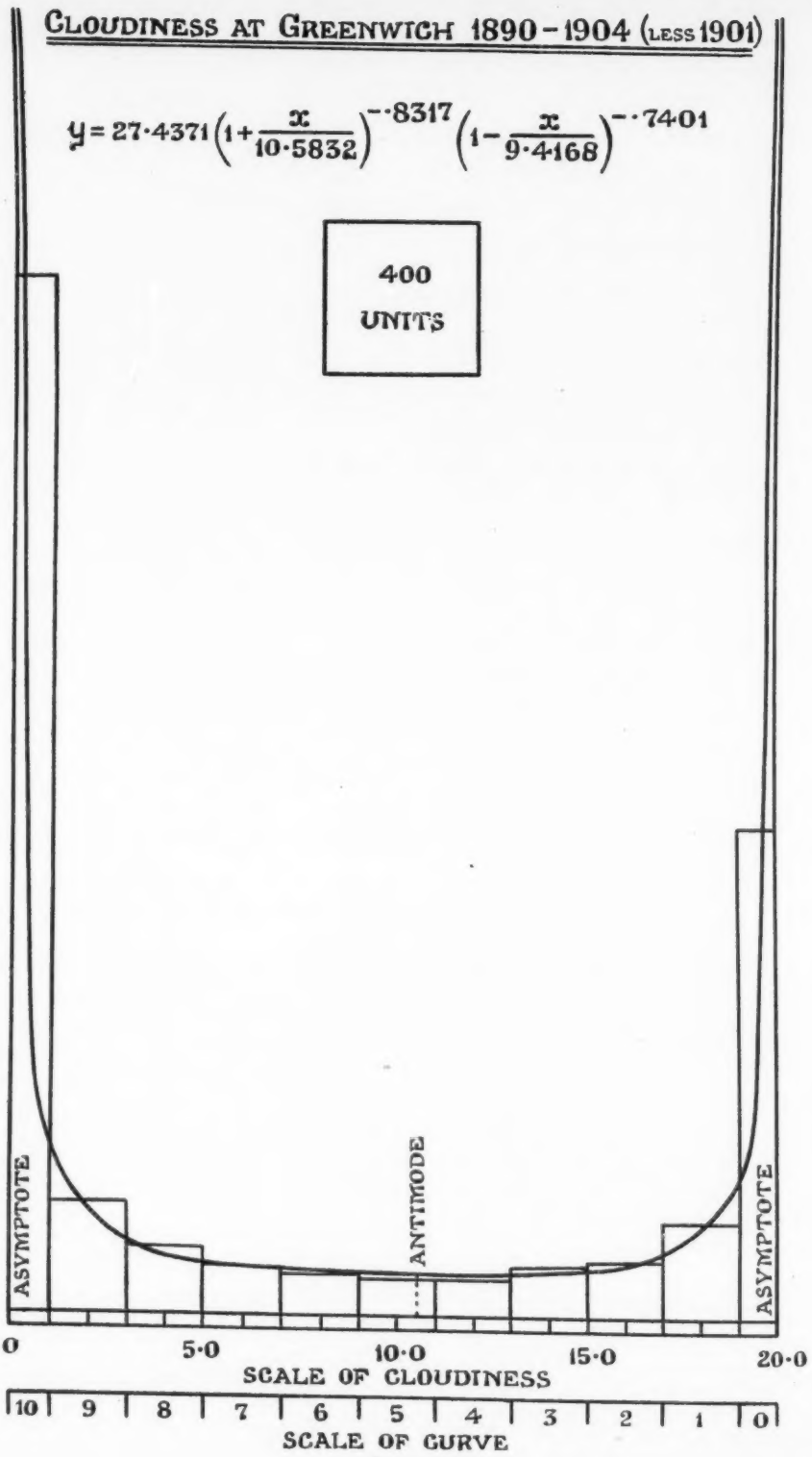

CLOUDINESS AT GREENWICH 1890-1904 (LESS 1901)
$y = 27{\cdot}4371\left(1+\frac{x}{10{\cdot}5832}\right)^{-{\cdot}8317}\left(1-\frac{x}{9{\cdot}4168}\right)^{-{\cdot}7401}$
400
UNITS
ASYMPTOTE
ANTIMODE
ASYMPTOTE
0
5·0
10·0
15·0
20·0
SCALE OF CLOUDINESS
10 9 8 7 6 5 4 3 2 1 0
SCALE OF CURVE

Diagram VII.

True Area	Calculated Area	$\frac{(\text{Difference})^2}{\text{Calculated Value}}$	True Area	Calculated Area	$\frac{(\text{Difference})^2}{\text{Calculated Value}}$
676·0	678·0	·0059	22·5	27·5	·9091
86·1	83·2	·1011	22·5	27·7	·9762
61·9	56·4	·5363	22·5	28·3	1·1887
49·3	44·9	·4312	33·1	29·7	·1939
40·7	38·3	·1504	34·9	31·7	·3230
34·6	34·1	·0073	35·5	34·9	·0103
30·4	31·4	·0318	38·5	40·2	·0719
27·9	29·5	·0868	48·9	49·7	·0129
27·1	28·3	·0509	80·1	71·8	·9567
22·5	27·7	·9762	320·0	321·7	·0090
		Totals	1715	1715	$\chi^2 = 7{\cdot}0296$, $n' = 20$

Testing for Goodness of Fit, we have $P = {\cdot}994$, a remarkably good fit.

The corresponding values for the eleven subranges originally given are:

True Area	Calculated Area	$\frac{(\text{Difference})^2}{\text{Calculated Value}}$	True Area	Calculated Area	$\frac{(\text{Difference})^2}{\text{Calculated Value}}$
676	678·0	·0059	45	56·0	2·1607
148	139·6	·5298	68	61·4	·7094
90	83·2	·5558	74	75·1	·0161
65	65·5	·0038	129	121·5	·4630
55	57·8	·1356	320	321·7	·0090
45	55·2	1·8848			
		Totals	1715	1715	$\chi^2 = 6{\cdot}4739$, $n' = 11$

Testing for Goodness of Fit, $P = {\cdot}773$.

If we consider this illustration merely as a test of the value of our tables for cases of asymptotic frequency, we see that there is a remarkably good fit ($P = {\cdot}994$) to the frequencies from which the moment coefficients are calculated.

This is natural since we used our auxiliary curve to subdivide the subranges. The actual fit to the frequencies originally given ($P = {\cdot}773$) may still be considered very good for the irregular data used. We note that the fit is least satisfactory in the middle of the range, which is what we should expect from our method of splitting up the frequencies.

This suggests the following method for dealing with frequency distributions for degrees of cloudiness.

Consider the two terminal frequencies (no cloudiness and 10 degrees of cloudiness) as of range $\frac{1}{2}$ centred at $\frac{1}{4}$ and $9\frac{3}{4}$ respectively. Split up the remaining nine frequencies, centred at 1, 2, 3, etc. of range unity, into frequencies of range a half centred at $\frac{3}{4}$, $1\frac{1}{4}$, $1\frac{3}{4}$, etc. by the use of an auxiliary curve of the type

$$Z = N[1 + x^q(A + Bx + Cx^2 + Dx^3 + Ex^4)],$$

or possibly $Z = N[1 + x^q(A + Bx + Cx^2)]$ would do equally well. Calculate the moments of this new distribution of twenty subranges by means of equations (IX), then fit a Pearson Type I-curve in the usual manner, taking the range b as fixed and equal to twenty ($h = \frac{1}{2}$), thus only using the first three moments. We see that in the illustration given, this gives a fit of $P = \cdot 773$ by the Goodness of Fit test.

It may be worth while seeing what fit we obtain without the use of abruptness coefficients, but with a fixed range of ten for the eleven subranges. The constants are:

$$\mu_1' = 3{\cdot}939{,}359; \quad \mu_2' = 31{\cdot}310{,}350^-; \quad \mu_3' = 276{\cdot}003{,}827; \quad \mu_2 = 15{\cdot}791{,}804;$$
$$\mu_3 = 28{\cdot}241{,}757; \quad \beta_1 = {\cdot}202{,}529; \quad r = {\cdot}503{,}749; \quad e = {\cdot}060{,}261; \quad m_2 = -{\cdot}691{,}735^+;$$
$$m_1 = -{\cdot}804{,}516; \quad a_2 = 4{\cdot}623{,}122; \quad a_1 = 5{\cdot}376{,}878.$$

Distance, mean to mode = 1·496,292.

Mode is $x = 5{\cdot}435{,}651$.

Curve starts at + ·058,773 and ends at 10·058,773.

$y_0 = 61{\cdot}992{,}20$.

Equation of curve is

$$y = 61{\cdot}992{,}20\left(1 + \frac{x}{5{\cdot}376{,}878}\right)^{-{\cdot}804{,}516}\left(1 - \frac{x}{4{\cdot}623{,}122}\right)^{-{\cdot}691{,}735+},$$

whence we have the following table:

True Area	Calculated Area	$\frac{(\text{Difference})^2}{\text{Calculated Value}}$	True Area	Calculated Area	$\frac{(\text{Difference})^2}{\text{Calculated Value}}$
676	616·5	5·7425	45	62·7	4·9967
148	170·1	2·8713	68	67·5	·0037
90	97·1	·5192	74	80·1	·4645
65	75·6	1·4862	129	119·2	·8057
55	66·3	1·9259	320	297·4	1·7174
45	62·5	4·9000			
		Totals	1715	1715	$\chi^2 = 25{\cdot}4331$, $n' = 11$

Testing for Goodness of Fit, $P = {\cdot}0046$—a very poor result as we should expect. This demonstrates how important it is to use abruptness coefficients.

The conclusion we draw from this paper is: that for all the six examples we have tried the following method is successful. Select q by the use of Table I, calculate the crude moments about the end of the first subrange for all the frequencies except that on this first subrange, then calculate the K's from Table IV. The corrected moments about the end of the first subrange are then given by equations (VIII). If by any chance the value of $m_1 + 1$ in the Pearson Type curve does not approximately agree with q, then try the value for q which is nearest to the $m_1 + 1$ found. In all cases fix the start of the curve and use three moments.

My thanks are due to Professor Pearson who first drew my attention to the subject of this investigation, and whose kindly criticism and suggestions for illustrations have been most helpful throughout.

Note I. A paper by Paul H. Rider on the "Analysis of a U-shaped Frequency Distribution" was published in June, 1927, in the *Journal of the American Statistical Association.* In this an attempt is made to fit a curve to the cloudiness data from Breslau by superposing two Charlier curves. The result is shown graphically in Diagram VIII, p. 348. We notice that the data have a peculiar dip at 9 degrees which the author agrees may be due to faulty observation. By taking six constants instead of the four used by Pearson, a curve has been obtained which follows the data more closely—in fact it follows all irregularities of the data histogram and even exaggerates them. We can only say that by taking eleven constants we would undertake to follow the data even more closely, but our object being a *graduation* which might reasonably represent the data when freed from observational errors, we prefer to smooth out such bends. We have not yet come across any natural law which is represented by such an irregular graph, and further if we use merely ordinates at central values we see no reason for treating cloudiness as *discontinuous* material. Again, if we test Mr Rider's results for Goodness of Fit, though his χ^2 is a great deal lower than Pearson's, his probability is still zero to six places of decimals. It does not really seem to be very profitable to try to fit a curve to such irregular data.

Note II. In the paper by Eleanor Pairman and Karl Pearson, to which reference has been already made, it is noted that in many cases "we have details for the ends of the range on different subranges to those for the bulk of the curve" and that "we shall get more accurate corrective terms if they are used." The cases mentioned were deaths in infancy, disease incidence in infancy, wages, incomes and house-valuations. The only one of these cases which has so far been found to give any possibility of an asymptotic curve is infant mortality, and then only for five causes of death, the others proving on examination to be very abrupt but not asymptotic. These five causes were Syphilis, Congenital Malformations, Congenital Debility, Premature Birth and Injury at Birth. In the first three the frequencies of death began to increase at the end of the first year of life; and it was obviously impossible to fit a curve to deaths from Premature Birth unless the degree of prematurity were known. An attempt was made to fit a curve to the American data for deaths from Injury at Birth, but great difficulties were experienced as they were given partly in day, partly in week and partly in month units, and also possessed some obvious irregularities; no satisfactory result has yet been obtained.

A method for taking advantage of such auxiliary subranges has however been considered, and various tables have been computed; but as it has not yet been found possible to verify them by practical examples they are not given here. They are, however, available for any future investigator.

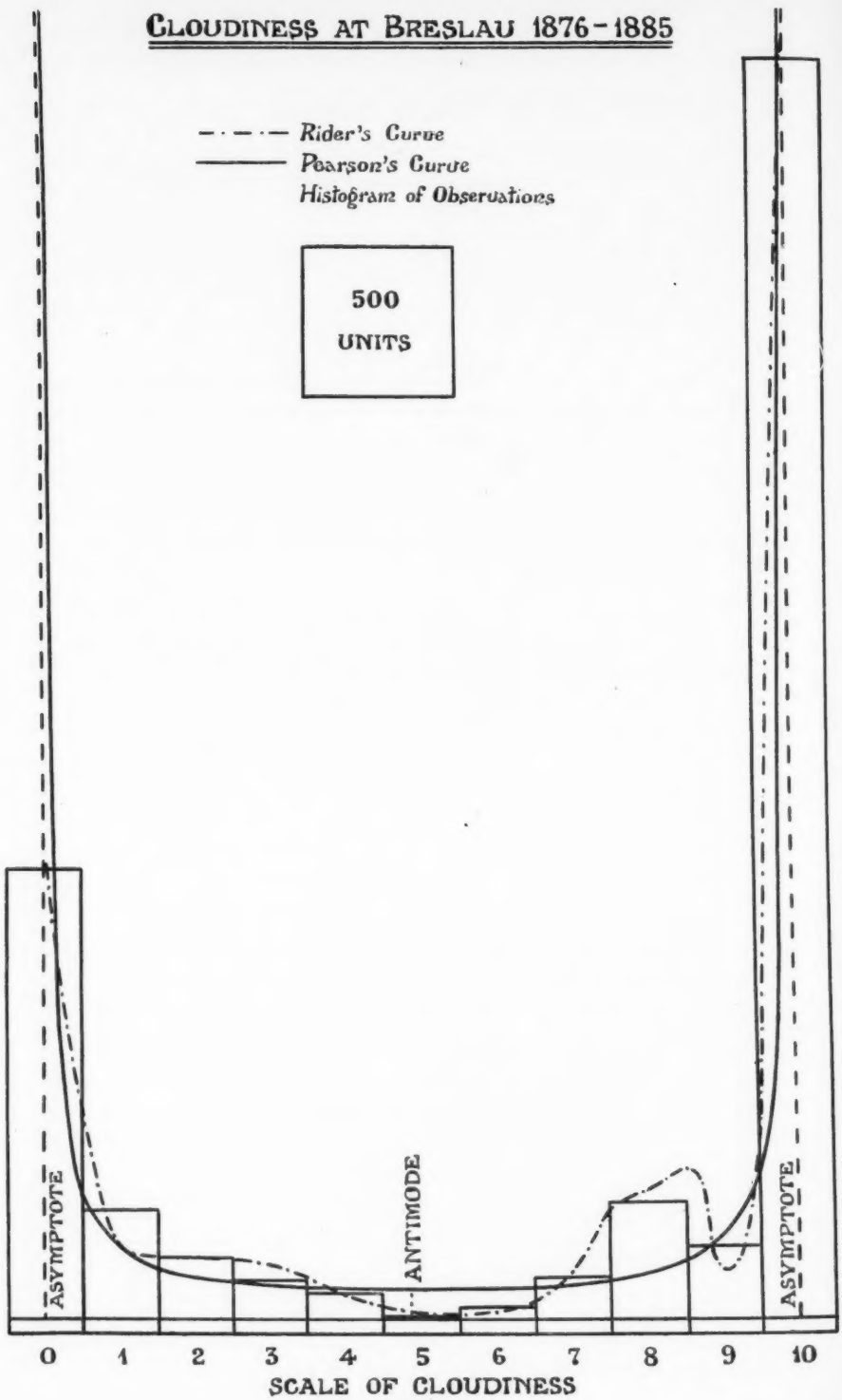


Diagram VIII.

TABLE I.

To select "q."

Values of n_6'.

q	$+n_5'$	$-n_4'$	$+n_3'$	$-n_2'$	$+n_1'$
·00	4·000,00	6·000,00	4·000,00	1·000,00	0·000,00
·01	4·009,12	6·031,50	4·038,05+	1·017,18	·000,90
·02	4·018,27	6·063,16	4·076,44	1·034,64	·001,85−
·03	4·027,42	6·094,96	4·115,16	1·052,38	·002,85−
·04	4·036,60	6·126,91	4·154,23	1·070,39	·003,91
·05	4·045,79	6·159,01	4·193,64	1·088,70	·005,02
·06	4·055,00−	6·191,27	4·233,39	1·107,30	·006,20
·07	4·064,22	6·223,67	4·273,50−	1·126,19	·007,44
·08	4·073,46	6·256,23	4·313,95+	1·145,38	·008,74
·09	4·082,72	6·288,94	4·354,76	1·164,88	·010,11
·1	4·092,00	6·321,80	4·395,93	1·184,68	·011,55−
·2	4·185,69	6·659,03	4·827,95+	1·400,70	·030,27
·3	4·281,10	7·012,37	5·299,08	1·652,87	·058,90
·4	4·378,27	7·382,52	5·812,56	1·946,67	·101,00
·5	4·477,23	7·770,22	6·371,91	2·288,34	·161,15−
·6	4·478,00	8·176,24	6·980,93	2·684,99	·245,17
·7	4·680,63	8·601,38	7·643,67	3·144,68	·360,47
·8	4·785,16	9·046,46	8·364,55−	3·676,58	·516,39
·9	4·891,60	9·512,37	9·148,29	4·291,09	·724,67
1·0	5·000,00	10·000,00	10·000,00	5·000,00	1·000,00

TABLE II.

$n_1'\mu_1''$.

q	$-n_5'$	$+n_4'$	$-n_3'$	$+n_2'$	$-n_1'$
·00	·348,611,111	1·420,833,333	2·212,500,000	1·640,277,778	1·000,000,000
·01	·337,052,155+	1·377,748,898	2·154,269,510	1·608,085,274	·995,720,335−
·02	·325,920,901	1·336,150,029	2·097,842,168	1·576,701,212	·991,464,828
·03	·315,199,524	1·295,977,975+	2·043,151,833	1·546,099,611	·987,233,322
·04	·304,871,044	1·257,176,655+	1·990,135,231	1·516,255,532	·983,025,661
·05	·294,919,274	1·219,692,523	1·938,731,809	1·487,145,029	·978,841,691
·06	·285,328,782	1·183,474,435−	1·888,883,596	1·458,745,101	·974,681,257
·07	·276,084,849	1·148,473,524	1·840,535,076	1·431,033,645−	·970,544,209
·08	·267,173,435−	1·114,643,089	1·793,633,065−	1·403,989,417	·966,430,394
·09	·258,581,138	1·081,938,478	1·748,126,600	1·377,591,989	·962,339,662
·1	·250,295,166	1·050,316,992	1·703,966,827	1·351,821,712	·958,271,865+
·2	·181,846,692	·785,494,654	1·327,010,619	1·124,779,734	·918,823,364
·3	·133,473,847	·593,351,364	1·043,358,325+	·943,454,131	·881,512,010
·4	·098,847,171	·452,136,857	·827,163,395−	·796,849,912	·846,205,880
·5	·073,782,714	·347,188,907	·660,549,745+	·677,069,086	·812,781,803
·6	·055,461,552	·268,428,056	·530,900,040	·578,314,290	·781,125,134
·7	·041,952,952	·208,807,221	·429,151,209	·496,250,299	·751,128,978
·8	·031,915,415+	·163,326,298	·348,692,994	·427,582,408	·722,693,605+
·9	·024,404,997	·128,391,269	·284,638,888	·369,770,901	·695,725,905−
1·0	·018,750,000	·101,388,889	·233,333,333	·320,833,333	·670,138,889

TABLE II (*continued*).

$n_1'\mu_2''$.

q	$+n_5'$	$-n_4'$	$+n_3'$	$-n_2'$	$+n_1'$
·00	·509,722,222	2·068,055,556	3·190,277,778	2·298,611,111	1·000,000,000
·01	·491,437,690	1·999,661,471	3·097,350,764	2·246,736,419	·992,840,159
·02	·473,876,359	1·933,800,012	3·007,538,094	2·196,296,879	·985,741,618
·03	·457,006,444	1·870,365,381	2·920,718,901	2·147,243,645⁻	·978,703,782
·04	·440,797,722	1·809,256,787	2·836,777,756	2·099,529,893	·971,726,057
·05	·425,221,445⁺	1·750,378,175⁺	2·755,604,379	2·053,110,720	·964,807,860
·06	·410,250,258	1·693,637,976	2·677,093,379	2·007,943,052	·957,948,613
·07	·395,858,125⁻	1·638,948,868	2·601,144,002	1·963,985,555⁻	·951,147,744
·08	·382,020,257	1·586,227,555⁺	2·527,659,900	1·921,198,548	·944,404,687
·09	·368,713,047	1·535,394,560	2·456,548,902	1·879,543,929	·937,718,883
·1	·355,914,005⁻	1·486,374,019	2·387,722,815⁻	1·838,985,097	·931,089,778
·2	·251,616,015⁻	1·081,409,527	1·808,207,172	1·486,556,641	·867,799,632
·3	·179,805,027	·795,134,092	1·383,315,321	1·212,311,406	·809,612,958
·4	·129,708,367	·590,084,242	1·067,691,574	·996,220,560	·756,065,739
·5	·094,357,154	·441,521,782	·830,558,572	·824,113,680	·706,740,651
·6	·069,157,670	·332,794,025⁺	·650,612,131	·685,753,113	·661,261,889
·7	·051,032,190	·252,502,977	·512,849,951	·573,607,979	·619,290,643
·8	·037,889,121	·192,732,854	·406,547,787	·482,050,991	·580,521,082
·9	·028,289,036	·147,915,106	·323,937,439	·406,819,019	·544,676,846
1·0	·021,230,159	·114,087,302	·259,325,397	·344,642,857	·511,507,937

$n_1'\mu_3''$.

q	$-n_5'$	$+n_4'$	$-n_3'$	$+n_2'$	$-n_1'$
·00	·605,357,143	2·449,404,762	3·757,738,095⁺	2·663,690,476	1·000,000,000
·01	·582,394,737	2·363,294,726	3·640,294,889	2·597,670,256	·990,639,993
·02	·560,384,903	2·280,539,516	3·527,014,097	2·533,603,950⁺	·981,380,506
·03	·539,284,120	2·200,993,279	3·417,727,350⁻	2·471,422,128	·972,220,333
·04	·519,051,072	2·124,517,267	3·312,274,067	2·411,058,305⁺	·963,158,286
·05	·499,646,518	2·050,979,455⁻	3·210,501,048	2·352,448,797	·954,193,191
·06	·481,033,181	1·980,254,169	3·112,262,082	2·295,532,578	·945,323,891
·07	·463,175,633	1·912,221,748	3·017,417,588	2·240,251,153	·936,549,243
·08	·446,040,202	1·846,768,219	2·925,834,268	2·186,548,431	·927,868,119
·09	·429,594,868	1·783,784,996	2·837,384,791	2·134,370,613	·919,279,408
·1	·413,809,176	1·723,168,594	2·751,947,482	2·083,666,078	·910,782,009
·2	·286,496,368	1·227,568,195⁺	2·039,995,560	1·647,661,301	·830,598,706
·3	·200,564,156	·884,100,493	1·528,228,154	1·315,019,404	·758,439,437
·4	·141,785,267	·642,875,541	1·155,459,749	1·058,012,378	·693,416,763
·5	·101,108,297	·471,474,675⁺	·880,790,011	·857,268,881	·634,750,209
·6	·072,666,589	·348,427,881	·676,338,042	·698,975,747	·581,752,328
·7	·052,595,818	·259,280,016	·522,773,395⁻	·573,108,001	·533,816,687
·8	·038,314,297	·194,157,347	·406,493,821	·472,277,467	·490,407,578
·9	·028,075,486	·146,228,827	·317,801,752	·390,965,397	·451,051,083
1·0	·020,684,524	·110,714,286	·249,702,381	·325,000,000	·415,327,381

TABLE II (*continued*).

$n_1' \mu_4''$.

q	$+n_5'$	$-n_4'$	$+n_3'$	$-n_2'$	$+n_1'$
·00	·669,246,032	2·702,976,190	4·131,547,619	2·897,817,460	1·000,000,000
·01	·642,741,032	2·603,384,896	3·995,314,950⁻	2·820,816,125⁺	·988,853,670
·02	·617,377,098	2·507,829,578	3·864,123,682	2·746,214,176	·977,846,549
·03	·593,100,484	2·416,129,154	3·737,763,098	2·673,923,356	·966,976,723
·04	·569,860,225⁻	2·328,111,578	3·616,032,461	2·603,859,235⁺	·956,242,310
·05	·547,607,985⁻	2·243,613,341	3·498,740,475⁺	2·535,941,014	·945,641,453
·06	·526,297,902	2·162,478,996	3·385,704,789	2·470,091,342	·935,172,325⁺
·07	·505,886,451	2·084,560,716	3·276,751,518	2·406,236,148	·924,833,125⁻
·08	·486,332,315⁻	2·009,717,880	3·171,714,803	2·344,304,475⁺	·914,622,078
·09	·467,596,255⁻	1·937,816,682	3·070,436,388	2·284,228,330	·904,537,438
·1	·449,641,002	1·868,729,761	2·972,765,231	2·225,942,537	·894,577,482
·2	·306,041,101	1·308,603,282	2·165,697,799	1·728,979,624	·801,474,553
·3	·210,673,314	·926,641,096	1·594,834,843	1·355,881,989	·719,123,762
·4	·146,480,974	·662,652,275⁻	1·185,629,153	1·072,196,971	·646,172,548
·5	·102,760,653	·478,040,680	·888,868,775⁺	·854,114,913	·581,451,699
·6	·072,670,572	·347,586,985⁻	·671,432,205⁺	·684,846,886	·523,948,751
·7	·051,766,689	·254,541,071	·510,650,857	·552,347,524	·472,785,494
·8	·037,121,452	·187,617,231	·390,780,891	·447,843,394	·427,198,921
·9	·026,782,049	·139,114,259	·300,745,049	·364,858,945⁻	·386,525,054
1·0	·019,431,217	·103,716,931	·232,658,730	·298,558,201	·350,185,185⁺

TABLE III.

a_1.

q	$+n_5'$	$-n_4'$	$+n_3'$	$-n_2'$	$-n_1'$
·00	·250,000,000	1·083,333,333	1·916,666,667	2·083,333,333	0·000,000,000
·01	·246,008,611	1·068,968,328	1·898,253,684	2·074,116,298	·000,782,964
·02	·242,080,946	1·054,792,317	1·880,009,840	2·064,920,978	·001,587,644
·03	·238,215,989	1·040,802,836	1·861,933,685⁻	2·055,747,506	·002,414,172
·04	·234,412,739	1·026,997,457	1·844,023,777	2·046,596,013	·003,262,680
·05	·230,670,209	1·013,373,780	1·826,278,688	2·037,466,628	·004,133,295⁻
·06	·226,987,430	·999,929,437	1·808,697,000	2·028,359,478	·005,026,144
·07	·223,363,450⁻	·986,662,091	1·791,277,305⁺	2·019,274,687	·005,941,353
·08	·219,797,328	·973,569,433	1·774,018,209	2·010,212,378	·006,879,045⁻
·09	·216,288,141	·960,649,187	1·756,918,324	2·001,172,673	·007,839,340
·1	·212,834,981	·947,899,104	1·739,976,276	1·992,155,690	·008,822,357
·2	·181,194,916	·829,282,795⁺	1·578,941,890	1·903,260,363	·019,927,030
·3	·154,258,466	·725,413,475⁻	1·432,255,805⁺	1·816,753,780	·033,420,447
·4	·131,326,390	·634,472,513	1·298,709,532	1·732,723,601	·049,390,268
·5	·111,803,399	·554,863,268	1·177,187,540	1·651,239,585⁻	·067,906,252
·6	·095,182,697	·485,184,345⁺	1·066,661,229	1·572,354,593	·089,021,260
·7	·081,032,830	·424,206,026	·966,183,145⁻	1·496,105,682	·112,772,349
·8	·068,986,483	·370,849,487	·874,881,452	1·422,515,258	·139,181,924
·9	·058,730,947	·324,168,505⁻	·791,954,670	1·351,592,256	·168,258,922
1·0	·050,000,000	·283,333,333	·716,666,667	1·283,333,333	·200,000,000

TABLE III (*continued*).

a_2.

q	$-n_5'$	$+n_4'$	$-n_3'$	n_2'	$+n_1'$
·00	·916,666,667	3·750,000,000	5·750,000,000	+2·916,666,667	·000,000,000
·01	·897,111,401	3·679,008,348	5·657,850,251	+2·869,503,976	·004,403,976
·02	·877,946,899	3·609,234,592	5·566,913,486	+2·822,639,386	·008,906,053
·03	·859,165,668	3·540,659,092	5·477,175,700	+2·776,072,675⁻	·013,506,008
·04	·840,760,355⁺	3·473,262,514	5·388,623,028	+2·729,803,599	·018,203,599
·05	·822,723,744	3·407,025,817	5·301,241,751	+2·683,831,902	·022,998,568
·06	·805,048,753	3·341,930,259	5·215,018,285⁺	+2·638,157,305⁻	·027,890,638
·07	·787,728,433	3·277,957,383	5·129,939,189	+2·592,779,515⁻	·032,879,515⁻
·08	·770,755,963	3·215,089,019	5·045,991,157	+2·547,698,220	·037,964,887
·09	·754,124,652	3·153,307,278	4·963,161,021	+2·502,913,093	·043,146,426
·1	·737,827,933	3·092,594,546	4·881,435,747	+2·458,423,787	·048,423,787
·2	·591,903,392	2·540,577,512	4·122,177,451	+2·029,713,197	·106,379,863
·3	·473,059,295⁻	2·077,989,333	3·460,028,486	+1·630,086,531	·173,419,865⁻
·4	·376,468,985⁺	1·691,188,054	2·884,009,453	+1·258,949,162	·248,949,162
·5	·298,142,397	1·368,524,270	2·384,252,480	+ ·915,579,166	·332,245,832
·6	·234,783,986	1·100,060,211	1·951,902,751	+ ·599,145,877	·422,479,210
·7	·183,674,414	·877,327,182	1·579,027,019	+ ·308,727,749	·518,727,749
·8	·142,572,065⁺	·693,116,279	1·258,528,860	+ ·043,329,275⁺	·619,995,942
·9	·109,631,101	·541,297,966	·984,070,371	− ·198,103,166	·725,230,168
1·0	·083,333,333	·416,666,667	·750,000,000	− ·416,666,667	·833,333,333

a_3.

q	n_5'	n_4'	n_3'	n_2'	$-n_1'$
·00	+1·500,000,000	−5·500,000,000	+6·500,000,000	−2·500,000,000	·000,000,000
·01	+1·441,684,262	−5·284,817,276	+6·214,058,189	−2·347,590,832	·016,666,832
·02	+1·384,993,511	−5·074,991,588	+5·934,038,264	−2·197,253,138	·033,561,138
·03	+1·329,888,404	−4·870,411,158	+5·659,846,120	−2·048,975,159	·050,677,159
·04	+1·276,330,479	−4·670,966,312	+5·391,388,909	−1·902,745,073	·068,009,073
·05	+1·224,282,132	−4·476,549,446	+5·128,575,027	−1·758,550,999	·085,550,999
·06	+1·173,706,605⁻	−4·287,054,989	+4·871,314,097	−1·616,380,995⁺	·103,296,995⁺
·07	+1·124,567,960	−4·102,379,367	+4·619,516,955⁺	−1·476,223,067	·121,241,067
·08	+1·076,831,069	−3·922,420,967	+4·373,095,641	−1·338,065,161	·139,377,161
·09	+1·030,461,591	−3·747,080,106	+4·131,963,378	−1·201,895,174	·157,699,174
·10	+ ·985,425,960	−3·576,258,991	+3·896,034,564	−1·067,700,951	·176,200,951
·20	+ ·601,567,121	−2·096,408,368	+1·804,915,623	+ ·168,280,564	·369,719,430
·30	+ ·319,315,024	− ·973,802,729	+ ·148,185,297	+1·220,358,460	·574,141,540
·40	+ ·115,567,223	− ·137,146,415⁻	−1·142,401,078	+2·101,232,374	·782,767,626
·50	− ·027,950,850⁻	+ ·472,049,150⁺	−2·126,027,061	+2·823,720,407	·988,779,593
·60	− ·125,641,160	+ ·901,608,504	−2·853,857,785⁻	+3·400,609,712	1·185,390,288
·70	− ·188,806,494	+1·190,495,582	−3·369,980,897	+3·844,525,366	1·365,974,634
·80	− ·226,275,664	+1·370,328,908	−3·712,253,327	+4·167,817,389	1·524,182,611
·90	− ·244,908,050⁻	+1·466,652,499	−3·913,061,200	+4·382,465,400	1·654,034,600
1·00	− ·250,000,000	+1·500,000,000	−4·000,000,000	+4·500,000,000	1·750,000,000

TABLE III (*continued*).

a_4.

q	n_5'	n_4'	n_3'	n_2'	$+n_1'$
·00	−1·000,000,000	+3·000,000,000	− 3·000,000,000	+ 1·000,000,000	·000,000,000
·01	− ·852,025,239	+2·439,751,274	− 2·227,973,805+	+ ·562,600,217	·058,931,207
·02	− ·710,160,918	+1·900,662,836	− 1·481,309,257	+ ·135,986,141	·118,033,981
·03	− ·574,217,737	+1·382,159,227	− ·759,456,556	− ·279,963,720	·177,272,470
·04	− ·444,011,482	+ ·883,678,251	− ·061,875,366	− ·685,370,531	·236,610,909
·05	− ·319,362,904	+ ·404,670,698	+ ·611,965,328	− 1·080,355,121	·296,013,629
·06	− ·200,097,591	− ·055,399,924	+ 1·262,587,353	− 1·465,037,963	·355,445,077
·07	− ·086,045,855−	− ·497,057,650+	+ 1·890,503,478	− 1·839,539,171	·414,869,819
·08	+ ·022,957,391	− ·920,814,325+	+ 2·496,217,551	− 2·203,978,481	·474,252,559
·09	+ ·127,072,744	−1·327,169,850+	+ 3·080,224,620	− 2·558,475,241	·533,558,149
·10	+ ·226,456,419	−1·716,612,438	+ 3·643,011,069	− 2·903,148,400	·592,751,600
·20	+ ·991,498,580	−4·780,350,105+	+ 8·204,797,398	− 5·835,442,087	1·170,957,913
·30	+1·426,582,291	− 6·611,859,902	+11·119,866,240	− 7·912,831,911	1·701,068,089
·40	+1·630,548,461	− 7·540,659,205+	+12·750,019,538	− 9·246,059,275−	2·152,340,725+
·50	+1·677,050,983	− 7·822,949,017	+13·394,673,376	− 9·939,850,404	2·497,649,596
·60	+1·620,390,233	− 7·656,196,570	+13·299,926,060	−10·092,310,182	2·714,089,818
·70	+1·500,079,746	− 7·191,039,047	+12·666,499,084	− 9·794,499,593	2·783,400,407
·80	+1·344,408,582	− 6·540,970,693	+11·656,666,873	− 9·130,178,559	2·692,221,441
·90	+1·173,209,400	− 5·790,200,028	+10·400,282,735−	− 8·175,694,918	2·432,205,082
1·00	+1·000,000,000	− 5·000,000,000	+ 9·000,000,000	− 7·000,000,000	2·000,000,000

a_5.

q	$-n_5'$	$+n_4'$	$-n_3'$	$+n_2'$	$-n_1'$
·00	·000,000,000	·000,000,000	·000,000,000	·000,000,000	·000,000,000
·01	·445,496,981	1·729,231,540	2·467,592,290	1·481,479,543	·239,271,919
·02	·865,971,768	3·368,992,773	4·822,507,268	2·909,869,574	·476,973,830
·03	1·262,377,028	4·922,319,719	7·067,900,676	4·286,091,610	·712,936,877
·04	1·635,635,324	6·392,164,841	9·206,860,660	5·611,059,695+	·946,994,807
·05	1·986,639,964	7·781,399,054	11·242,408,959	6·885,680,352	1·178,984,023
·06	2·316,255,826	9·092,813,698	13·177,502,082	8·110,852,542	1·408,743,636
·07	2·625,320,164	10·329,122,460	15·015,032,460	9·287,467,631	1·636,115,512
·08	2·914,643,391	11·492,963,258	16·757,829,591	10·416,409,351	1·860,944,326
·09	3·185,009,843	12·586,900,087	18·408,661,152	11·498,553,767	2·083,077,601
·10	3·437,178,520	13·613,424,813	19·970,234,116	12·534,769,243	2·302,365,757
·20	5·108,247,070	20·671,060,293	31·230,998,122	20·556,313,787	4·308,006,213
·30	5·631,899,453	23·256,087,239	36·132,680,857	24·882,094,652	5·887,460,345−
·40	5·446,368,055+	22·912,696,933	36·536,517,313	26·250,525,738	6·943,714,262
·50	4·856,460,139	20·768,539,861	33·899,313,753	25·320,879,172	7·413,495,828
·60	4·069,250,659	17·631,833,783	29·345,635,431	22·674,644,444	7·267,115,556
·70	3·221,014,470	14·068,006,546	23·729,027,854	18·818,091,927	6·507,103,073
·80	2·397,418,260	10·459,866,823	17·683,686,565−	14·185,800,597	5·165,879,403
·90	1·648,548,321	7·054,564,765+	11·667,843,417	9·144,925,452	3·302,689,548
1·00	1·000,000,000	4·000,000,000	6·000,000,000	4·000,000,000	1·000,000,000

TABLE IV.

K_1.

q	$-n_5$	$+n_4$	$-n_3$	$+n_2$	$-n_1$
·00	·329,861,111	1·338,194,444	2·061,805,556	1·470,138,889	1·000,000,000
·01	·318,568,509	1·296,065,411	2·004,793,940	1·438,551,783	·995,770,346
·02	·307,699,727	1·255,410,677	1·949,575,875+	1·407,772,430	·991,566,292
·03	·297,237,004	1·216,171,641	1·896,085,317	1·377,774,854	·987,387,694
·04	·287,163,418	1·178,292,368	1·844,259,083	1·348,534,117	·983,234,410
·05	·277,462,844	1·141,719,460	1·794,036,712	1·320,026,276	·979,106,298
·06	·268,119,906	1·106,401,913	1·745,360,324	1·292,228,334	·975,003,220
·07	·259,119,945-	1·072,291,004	1·698,174,492	1·265,118,189	·970,925,036
·08	·250,448,973	1·039,340,168	1·652,426,119	1·238,674,601	·966,871,608
·09	·242,093,647	1·007,504,891	1·608,064,330	1·212,877,142	·962,842,799
·10	·234,041,228	·976,742,606	1·565,040,355-	1·187,706,165+	·958,838,474
·20	·167,751,549	·719,982,999	1·198,971,727	·966,620,753	·920,112,912
·30	·121,248,708	·535,021,796	·925,404,352	·791,185,861	·883,694,319
·40	·088,243,920	·400,212,324	·718,559,151	·650,405,973	·849,464,178
·50	·064,587,541	·300,981,467	·560,618,977	·536,381,284	·817,312,508
·60	·047,489,724	·227,326,856	·439,018,337	·443,311,494	·787,137,511
·70	·035,044,500-	·172,268,464	·344,740,108	·366,857,498	·758,844,668
·80	·025,931,549	·130,864,835-	·271,215,077	·303,719,942	·732,346,008
·90	·019,225,117	·099,573,495-	·213,593,699	·251,353,867	·707,559,427
1·00	·014,269,180	·075,826,720	·168,253,968	·207,771,164	·684,408,069

K_2.

q	$+n_5$	$-n_4$	$+n_3$	$-n_2$	$+n_1$
·00	·517,030,423	2·098,313,492	3·237,202,381	2·322,585,979	1·000,000,000
·01	·498,631,864	2·029,513,077	3·143,762,752	2·270,462,907	·992,822,947
·02	·480,957,741	1·963,248,440	3·053,439,188	2·219,773,905-	·985,706,434
·03	·463,976,271	1·899,413,810	2·966,110,889	2·170,470,164	·978,649,853
·04	·447,657,229	1·837,908,420	2·881,662,486	2·122,504,900	·971,652,604
·05	·431,971,867	1·778,636,237	2·799,983,763	2·075,833,246	·964,714,093
·06	·416,892,828	1·721,505,715-	2·720,969,387	2·030,412,167	·957,833,732
·07	·402,394,074	1·666,429,550+	2·644,518,662	1·986,200,364	·951,010,940
·08	·388,450,815-	1·613,324,466	2·570,535,295-	1·943,158,195+	·944,245,142
·09	·375,039,440	1·562,110,999	2·498,927,170	1·901,247,595-	·937,535,770
·10	·362,137,457	1·512,713,303	2·429,606,145+	1·860,431,998	·930,882,262
·20	·256,876,419	1·104,161,810	1·845,271,878	1·505,400,628	·867,300,354
·30	·184,218,941	·814,637,133	1·415,826,097	1·228,512,138	·808,730,315+
·40	·133,384,811	·606,671,080	1·095,941,262	1·009,769,362	·754,702,916
·50	·097,396,255-	·455,513,105-	·854,856,798	·835,030,494	·704,797,878
·60	·071,650,047	·344,493,005-	·671,276,111	·694,083,399	·658,638,746
·70	·053,058,869	·262,192,026	·530,197,167	·579,419,632	·615,888,348
·80	·039,521,801	·200,671,843	·420,890,245+	·485,431,307	·576,244,734
·90	·029,590,594	·154,340,671	·335,577,272	·407,871,762	·539,437,561
1·00	·022,255,291	·119,212,963	·268,551,587	·343,485,450-	·505,224,868

TABLE IV (*continued*).

K_3.

q	$-n_5$	$+n_4$	$-n_3$	$+n_2$	$-n_1$
·00	·608,630,952	2·465,575,397	3·792,757,937	2·710,813,492	1·000,000,000
·01	·585,638,062	2·379,353,058	3·675,164,710	2·644,710,924	·990,628,564
·02	·563,598,718	2·296,488,960	3·561,738,113	2·580,564,234	·981,357,720
·03	·542,469,355⁻	2·216,837,094	3·452,309,599	2·518,303,921	·972,186,265
·04	·522,208,610	2·140,258,563	3·346,718,413	2·457,863,432	·963,113,013
·05	·502,777,201	2·066,621,194	3·244,811,187	2·399,179,016	·954,136,792
·06	·484,137,807	1·995,799,175⁺	3·146,441,545⁺	2·342,189,579	·945,256,448
·07	·466,254,963	1·927,672,708	3·051,469,746	2·286,836,561	·936,470,838
·08	·449,094,957	1·862,127,690	2·959,762,339	2·233,063,807	·927,778,837
·09	·432,625,733	1·799,055,407	2·871,191,840	2·180,817,451	·919,179,332
·10	·416,816,800	1·738,352,255⁻	2·785,636,429	2·130,045,811	·910,671,226
·20	·289,300,542	1·241,987,489	2·072,634,697	1·693,435,417	·830,385,350⁺
·30	·203,200,400	·897,881,173	1·559,976,710	1·260,267,710	·758,129,818
·40	·144,271,796	·656,078,498	1·186,388,260	1·102,765,151	·693,011,813
·50	·103,452,959	·484,119,473	·910,906,829	·901,514,899	·634,242,178
·60	·074,871,564	·360,509,746	·705,610,152	·742,669,912	·581,121,770
·70	·054,660,732	·270,781,844	·551,142,670	·616,178,452	·533,029,822
·80	·040,238,187	·205,057,921	·433,889,389	·514,632,139	·489,414,089
·90	·029,857,965⁻	·156,508,214	·344,149,229	·432,497,975⁺	·449,782,394
1·00	·022,326,389	·120,357,143	·274,927,556	·365,595,238	·413,695,437

K_4.

q	$+n_5$	$-n_4$	$+n_3$	$-n_2$	$+n_1$
·00	·662,665,344	2·675,297,619	4·087,996,032	2·875,363,757	1·000,000,000
·01	·636,212,784	2·575,880,689	3·951,958,580	2·798,432,979	·988,847,698
·02	·610,902,433	2·480,504,764	3·820,970,516	2·723,906,711	·977,835,264
·03	·586,680,471	2·388,988,518	3·694,820,850⁺	2·651,696,614	·966,960,808
·04	·563,495,865⁻	2·301,159,668	3·573,308,580	2·581,718,176	·956,222,470
·05	·541,300,211	2·216,854,475⁻	3·456,242,152	2·513,890,514	·945,618,417
·06	·520,047,582	2·135,917,267	3·343,438,958	2·448,136,198	·935,146,843
·07	·499,694,393	2·058,200,002	3·234,724,866	2·384,381,075⁺	·924,805,969
·08	·480,199,261	1·983,561,846	3·129,933,770	2·322,554,108	·914,594,045
·09	·461,522,894	1·911,868,788	3·028,907,176	2·262,587,222	·904,509,343
·10	·443,627,963	1·842,993,268	2·931,493,809	2·204,415,162	·894,550,165⁻
·20	·300,654,914	1·285,120,281	2·127,284,329	1·708,818,399	·801,505,673
·30	·205,928,193	·905,557,548	1·559,652,170	1·337,449,772	·719,318,813
·40	·142,360,800	·643,993,579	1·153,886,009	1·055,784,436	·646,653,654
·50	·099,229,825⁺	·461,746,773	·860,644,280	·839,945,738	·582,354,092
·60	·069,681,936	·333,539,545⁻	·646,704,854	·673,083,211	·525,416,976
·70	·049,267,234	·242,584,380	·489,322,717	·543,095,568	·474,969,464
·80	·035,056,311	·177,573,968	·372,697,659	·441,159,109	·430,249,925⁺
·90	·025,097,232	·130,797,271	·285,712,548	·360,753,626	·390,591,817
1·00	·018,075,397	·096,937,831	·220,456,349	·297,003,968	·355,410,053

THE DISTRIBUTION OF FREQUENCY CONSTANTS IN SMALL SAMPLES FROM SYMMETRICAL POPULATIONS.

(*Preliminary Notice.*)

By EGON S. PEARSON, D.Sc., assisted by N. K. ADYANTHĀYA.

It is hoped that the following preliminary results may be of some interest to practical workers, both in showing the extent to which the tests developed for the case of samples from a normal population are valid, and also in suggesting modifications of these tests when required. The range of symmetrical populations from $\beta_2 = 1{\cdot}8$ to $7{\cdot}1$ has been covered by adding to the theoretical values for the Rectangular and Normal Populations three sets of experimental sampling results as indicated below.

Population Curve	β_2 of grouped distribution	Samples	S.D. of population in terms of grouping unit
Type II	2·500	1000 of 2, 500 of 5, 500 of 10	6·3249
		500 of 20	10·5404
Type VII	4·122	1000 of 2, 500 of 5, 500 of 10, 500 of 20	5·6672
Type VII	7·069	1000 of 2, 1000 of 5, 500 of 10, 500 of 20	6·4482

The sampling was carried out with the help of Tippett's Random Numbers*, the population frequencies being obtained from "Student's" Tables of t†.

Suppose the variates of a sample of n to be arranged in order of magnitude, and u to be the highest and v the lowest value. Then the *Range* is $u - v$ and the *Centre* is $\frac{1}{2}(u+v)$. If n be odd the *Median* is the $\frac{1}{2}(n+1)$th observation, and if even is taken as the mid-point between the $(\frac{1}{2}n)$th and $(\frac{1}{2}n+1)$th. In Tables I and II theoretical values are given in italics; these were obtained as follows:—

Rectangular Population. $\beta_2 = 1{\cdot}8$.

Range; from formulae in *Biometrika*, Vol. XX^A. p. 217.

Centre, G; from equation (xl) of the same paper it follows that

$$\sigma_G = \sigma \sqrt{\frac{6}{(n+1)(n+2)}}.$$

Median; it is found that, n odd $\sigma_M = \sigma \sqrt{\frac{3}{n+2}}$,

$$n \text{ even } \sigma_M = \sigma \sqrt{\frac{3n}{(n+1)(n+2)}}.$$

* *Tracts for Computers*, No. xv.

† *Metron*, Vol. v. p. 114.

Normal Population $\beta_2 = 3{\cdot}0$.

Range; from Table VIII, *Biometrika*, Vol. XVIII. p. 192.

Centre; $\sigma_G = \sigma_u \sqrt{\frac{1}{2}(1 + r_{uv})}$, where σ_u and r_{uv} have been given by Tippett, *Biometrika*, Vol. XVII. p. 364 *et seq.*

Median; obtained up to $n = 5$ only, from the integrals computed and given on p. 178, *Biometrika*, Vol. XVIII.

The figures obtained experimentally are of course subject to errors of sampling, but if the whole series of values of mean and standard error of range and of standard error of centre are plotted either to n or to β_2, it will be seen that they change comparatively smoothly so that an interpolation adequate for many practical purposes can be obtained for any value of n between 2 and 20 or of β_2 between 1·8 and 7·1. As the definition of the median is different according to whether n be odd or even, interpolation for n is not in this case easy. The results may be considered briefly in order.

1. *The Distribution of Range.*

The mean and standard errors of Table I are given in terms of the population standard deviation. It will be seen that the mean changes very little with β_2 for samples of 10 or less, but that at 20 there is a somewhat greater change. The standard error of range changes however very considerably with β_2, increasing steadily as the population becomes more leptokurtic*. No useful interpolation can be made from the values of β_1 and β_2 for range, but the following points are of interest. If we define as "positively skew" a sampling curve with its longer tail towards increasing range, then, (*a*) for the Rectangular Population, the range distribution at $n = 2$ is a sloping straight line with positive skewness, at $n = 3$ it is a symmetrical platykurtic curve, while for $n > 3$ we have curves of increasing negative skewness; (*b*) at $\beta_2 = 2{\cdot}5$ the curves are platykurtic, almost symmetrical, but positively skew; (*c*) as the population β_2 increases the range curves for a given n become more and more positively skew; (*d*) for a given population and changing n, there is the same tendency as was observed in the normal case†, for the (β_1, β_2) point of the range curve to approach the normal point (0, 3) as n increases up to about 10, and then to recede.

For practical purposes it is necessary to know the chance of drawing a sample with a range greater than certain multiples of the population standard deviation. It would therefore be desirable to fit curves to the distribution of range; this has not yet been done, but the results given in Table II will give an adequate idea of the position. Here the permilles for the experimental sampling groups have been found by rough smoothing from the data; the rectangular results follow from equation (xxxix), *Biometrika*, Vol. XX^A. p. 210, while for the normal we are indebted to "Mathetes" who has lent the data used in computing the curves for

* At $\beta_2 = 4{\cdot}12$, chance sampling fluctuation appears to cause some irregularity at $n = 5$ or 10.

† See the diagram on p. 191, *Biometrika*, Vol. XVIII.

"Student's" paper on Routine Analysis*. The great length of the tails of the range curves obtained in sampling from leptokurtic populations will be seen at a glance. This is of considerable importance. "Student" has found, for example, that leptokurtic error systems are common in routine analysis*, and a value of $\beta_2 = 7{\cdot}0$ is probably not unduly exceptional. The analyst must decide therefore whether he should reject extreme observations as excessively improbable deviations on "normal theory," or accept them as perhaps rare but perfectly genuine variants in a leptokurtic system.

2. *The Distribution of Centre.*

The standard error of centre given in Table I is expressed as a multiple of $\sigma/\sqrt{n}$, the standard error of the mean. The centre is theoretically the best estimate of the mid-point of the rectangular population. The figures suggest that the standard error of the mean equals that of the centre for a population with $\beta_2 = 2{\cdot}2$ approximately, and that after this the ratio of the latter to the former steadily increases above unity. For samples of 5 or less, the centre is not so bad an estimate of the mid-point even when $\beta_2 = 4{\cdot}1$, but for larger samples it becomes very unsatisfactory.

3. *The Distribution of Median.*

Just as the centre is at its best as an estimate at $\beta_2 = 1{\cdot}8$, and the mean at $\beta_2 = 3{\cdot}0$, so the median begins to gain upon the latter as β_2 increases, and for the extreme sampling series, $\beta_2 = 7{\cdot}1$, we find that for $n = 10$ and 20, the standard error of median is less than that of mean. Further, in this case the sampling distribution of the median is possibly less leptokurtic than that of the mean, which if confirmed would be a point in favour of the former. We find

	$n=5$	10	20
β_2 of distribution of means (theory)	3·81	3·41	3·20
β_2 „ „ „ medians (experiment)	3·62 (·42)	3·13 (·27)	3·40 (·42)

where the standard errors are given in brackets. By using the median and taking account of the presence, but not of the exact value of the extreme variates, the large sampling errors arising from their inclusion on calculating the mean may be avoided. In problems, therefore, where it is known that the population is leptokurtic, the use of the median rather than the mean as a central estimate certainly appears to be worth consideration. As Table I shows, it is not a satisfactory estimate in the case of platykurtic distributions.

We are very grateful to Mr A. E. Stone for taking all but 400 of the 8000 samples involved.

* *Biometrika*, Vol. XIX. pp. 151—164.

TABLE I.

Distribution Constants.

Size of Sample.

β_2	—	2	3	4	5	10	20
1·80	Range: Mean	*1·155*	*1·732*	*2·078*	*2·309*	*2·834*	*3·134*
	Range: S.E.	*·816*	*·775*	*·693*	*·617*	*·386*	*·217*
	Range: β_1	*·320*	*·000*	*·082*	*·219*	*·773*	*1·265*
	Range: β_2	*2·400*	*2·143*	*2·357*	*2·625*	*3·648*	*4·569*
	Centre: S.E.	*1·000*	*·949*	*·894*	*·845*	*·674*	*·510*
	Median: S.E.	*1·000*	*1·342*	*1·265*	*1·464*	*1·508*	*1·612*
2·50	Range: Mean	1·158	—	—	2·343	3·024	3·604
	Range: S.E.	·850	—	—	·786	·713	·549
	Range: β_1	—	—	—	·110	·008	·039
	Range: β_2	—	—	—	2·869	2·884	2·705
	Centre: S.E.	*1·000*	—	—	1·062	1·101	1·266
	Median: S.E.	*1·000*	—	—	1·319	1·231	1·312
3·00	Range: Mean	*1·128*	*1·693*	*2·059*	*2·326*	*3·078*	*3·735*
	Range: S.E.	*·852*	*·888*	*·880*	*·864*	*·797*	*·729*
	Range: β_1	*·991*	*·417*	*·273*	*·217*	*·156*	*·161*
	Range: β_2	*3·869*	*3·286*	*3·188*	*3·169*	*3·22*	*3·26*
	Centre: S.E.	*1·000*	*1·042*	*1·092*	*1·142*	*1·362*	*1·691*
	Median: S.E.	*1·000*	*1·160*	*1·092*	*1·198*	—	*(→1·253)*
4·12	Range: Mean	1·118	—	—	2·360	3·085	3·961
	Range: S.E.	·890	—	—	1·010	·893	1·029
	Range: β_1	—	—	—	·614	·335	·603
	Range: β_2	—	—	—	4·748	3·377	3·693
	Centre: S.E.	*1·000*	—	—	1·182	1·569	2·411
	Median: S.E.	*1·000*	—	—	1·145	1·126	1·141
7·07	Range: Mean	1·020	—	—	2·245	3·034	3·905
	Range: S.E.	·892	—	—	1·100	1·229	1·300
	Range: β_1	—	—	—	9·042	2·228	2·633
	Range: β_2	—	—	—	19·280	6·898	7·970
	Centre: S.E.	*1·000*	—	—	1·372	2·015	2·947
	Median: S.E.	*1·000*	—	—	1·049	·979	·970

N.B. The Mean and Standard Error of Range are expressed in terms of the Population Standard Deviation, σ. The Standard Errors of Centre and Median are given as multiples of the Standard Error of the Mean, $\sigma/\sqrt{n}$. Figures in italics are theoretical, the others are obtained from experimental sampling.

TABLE II.

Permille of Samples with Range greater than Multiples of Population Standard Deviation.

Size of Sample	Population β_2	Range greater than 1σ	2σ	3σ	4σ	5σ	6σ	7σ	8σ	9σ	10σ	11σ
2	1·80	*506*	*179*	*18*	—	—	—	—	—	—	—	—
	2·50	501	174	27	3	—	—	—	—	—	—	—
	3·00	*480*	*157*	*34*	*5*	—	—	—	—	—	—	—
	4·12	468	154	36	8	3	1	—	—	—	—	—
	7·07	396	119	37	10	3	2	—	—	—	—	—
5	1·80	*973*	*701*	*136*	—	—	—	—	—	—	—	—
	2·50	968	646	190	34	—	—	—	—	—	—	—
	3·00	*956*	*619*	*211*	*38*	*4*	—	—	—	—	—	—
	4·12	922	609	240	62	8	4	2	2	—	—	—
	7·07	922	521	201	72	16	4	3	2	2	1	1
10	1·80	*1000*	*966*	*396*	—	—	—	—	—	—	—	—
	2·50	998	926	495	90	2	—	—	—	—	—	—
	3·00	*1000*	*922*	*515*	*127*	*15*	*1*	—	—	—	—	—
	4·12	1000	898	501	154	28	6	—	—	—	—	—
	7·07	994	831	436	154	68	30	16	4	2	2	—
20	1·80	*1000*	*1000*	*769*	—	—	—	—	—	—	—	—
	2·50	1000	1000	861	235	4	—	—	—	—	—	—
	3·00	*1000*	*998*	*856*	*349*	*55*	*4*	—	—	—	—	—
	4·12	1000	998	833	421	156	44	8	2	—	—	—
	7·07	1000	995	775	422	170	74	21	14	8	6	2

N.B. Figures in italics are theoretical, the others are obtained from experimental sampling.

In the case of $\beta_2 = 1{\cdot}8$ no sample range as great as 4σ is *possible*. For $\beta_2 = 3{\cdot}0$ the permilles beyond the last figure shown are less than unity; while for the other three cases no samples with range greater than the value indicated by the last figure in each row were *observed*.

ON THE MEANS OF SAMPLES FROM A U-SHAPED POPULATION.

By K. J. HOLZINGER, Ph.D. and A. E. R. CHURCH, M.A., D.Sc.

(1) If samples of N are taken from a population whose mean, second, third and fourth moment-coefficients, first two betas and standard deviation are respectively $\bar{\mu}_1'$, $\bar{\mu}_2$, $\bar{\mu}_3$, $\bar{\mu}_4$, $\bar{\beta}_1$, $\bar{\beta}_2$ and $\bar{\sigma}$, then the well-known formulae,

$$ {}_1M_1' = \bar{\mu}_1', \quad {}_1M_2 = \frac{\bar{\mu}_2}{N}, \quad {}_1M_3 = \frac{\bar{\mu}_3}{N^2}, \quad {}_1M_4 = \frac{\bar{\mu}_4}{N^3} + \frac{3(N-1)}{N^3} \cdot \bar{\mu}_2^2 ; $$

$$ {}_1B_1 = \frac{\bar{\beta}_1}{N}, \quad {}_1B_2 - 3 = \frac{\bar{\beta}_2 - 3}{N} ; \quad \Sigma_1 = \frac{\bar{\sigma}}{N^{\frac{1}{2}}}, $$

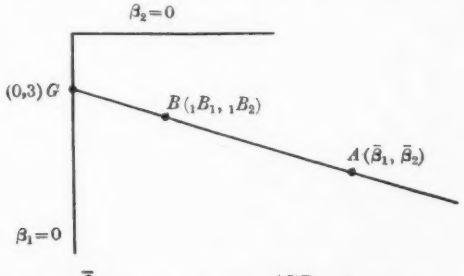
due to Tchebycheff, give the corresponding constants ${}_1M_1'$, ${}_1M_2$, ${}_1M_3$, ${}_1M_4$, ${}_1B_1$, ${}_1B_2$ and Σ_1 of the distribution of means in samples of N from that population, the sampling being on the basis of an "infinite" sampled population.

From these we have immediately

$$ \frac{\bar{\beta}_2 - 3}{\bar{\beta}_1} = \frac{{}_1B_2 - 3}{{}_1B_1} = \lambda, \text{ where } \lambda \text{ is independent of } N, $$

that is λ is a constant for all distributions of means of samples derived from the given population.

Thus taking rectangular axes $\beta_1 = 0$, $\beta_2 = 0$, the point $({}_1B_1, {}_1B_2)$ representing the betas of the distribution of means of samples of N, lies on the line ${}_1B_2 = \lambda \cdot {}_1B_1 + 3$, which passes through the Gaussian point (0, 3) and through the given point $(\bar{\beta}_1, \bar{\beta}_2)$ representing the betas of the original population.



Moreover, since ${}_1B_1 = \frac{\bar{\beta}_1}{N}$, in the diagram $\frac{GB}{GA} = \frac{1}{N}$ $(N \geqslant 1)$ and so B divides GA internally in the ratio of $1 : N$.

These considerations combined with the use of the Diagram, p. 66 of the *Tables for Statisticians*, and the use of the corresponding Pearson Curve, provide a possible

method of representing actual distributions of means in samples of N. This method of representation may fail, however, not only in experimental sampling but also in practical work owing to the effect of grouping. If a satisfactory method of sampling from a population which was actually represented by a smooth frequency curve, and not by a series of discrete frequency groups, could be devised, there is every probability that the discrepancies between the theoretical considerations of sampling and actual results would disappear, but at present no really satisfactory method of sampling under these conditions has been devised.

Of course, if a population represented by a continuous curve be divided into a very large number of frequency groups, the variety of possible samples to be obtained is very large and the means of such samples form a nearly continuous distribution and then practice agrees very fairly with theory. Now it has been shown* that when the sampled population approximates to an ordinary Pearson Curve as distinct from a U or J curve, the results of practical sampling (provided it is really random) agree quite well with theory, at any rate as regards means of samples; that is the distributions of means obtained are quite well represented by continuous curves even when the number of frequency groups in the sampled population is of the order 15 or 20. When the sampled population is represented by a U, J, or extremely leptokurtic curve, it is rather to be expected that the number of frequency groups into which it is divided for the purpose of sampling, will have to be considerably larger in order to obtain distributions of means of samples that are representable by continuous curves. This follows since the frequency tends to be concentrated into relatively few groups leaving a considerable number with little frequency in them. This is especially so in an extreme U-shaped distribution like the one selected here, and it has become apparent that the choice of eleven frequency groups was rather small.

This necessity of subdividing a sampled population into a limited number of frequency groups before actual sampling can be performed, is one of the greatest difficulties that arise in sampling, whilst, when the betas of a population are used in theoretical considerations, it must be remembered that, the population from which they are derived being a discontinuous one, their values will not be those of the original continuous curve.

There is therefore no *a priori* certainty that the method outlined above will give good representations of distributions of means of samples of N from an original population that consists of a U- or J-shaped frequency curve subdivided into a limited number of frequency groups. If, however, the method is applicable, starting with an approximate U-shaped population, the diagram above used in conjunction with *Tables for Statisticians*, XXXV, shows that, as N increases from 2, the form of the distribution of means of samples of N should be first U, then J, then the shape of an ordinary Pearson Type I curve and finally, when N is very large, normal in shape.

* "On the Means and Squared Standard Deviations of Small Samples from any Population." *Biometrika*, Vol. XVIII. pp. 321—394. Here also is given an explanation of the meaning of an "infinite" sampled population and a discussion on methods of sampling.

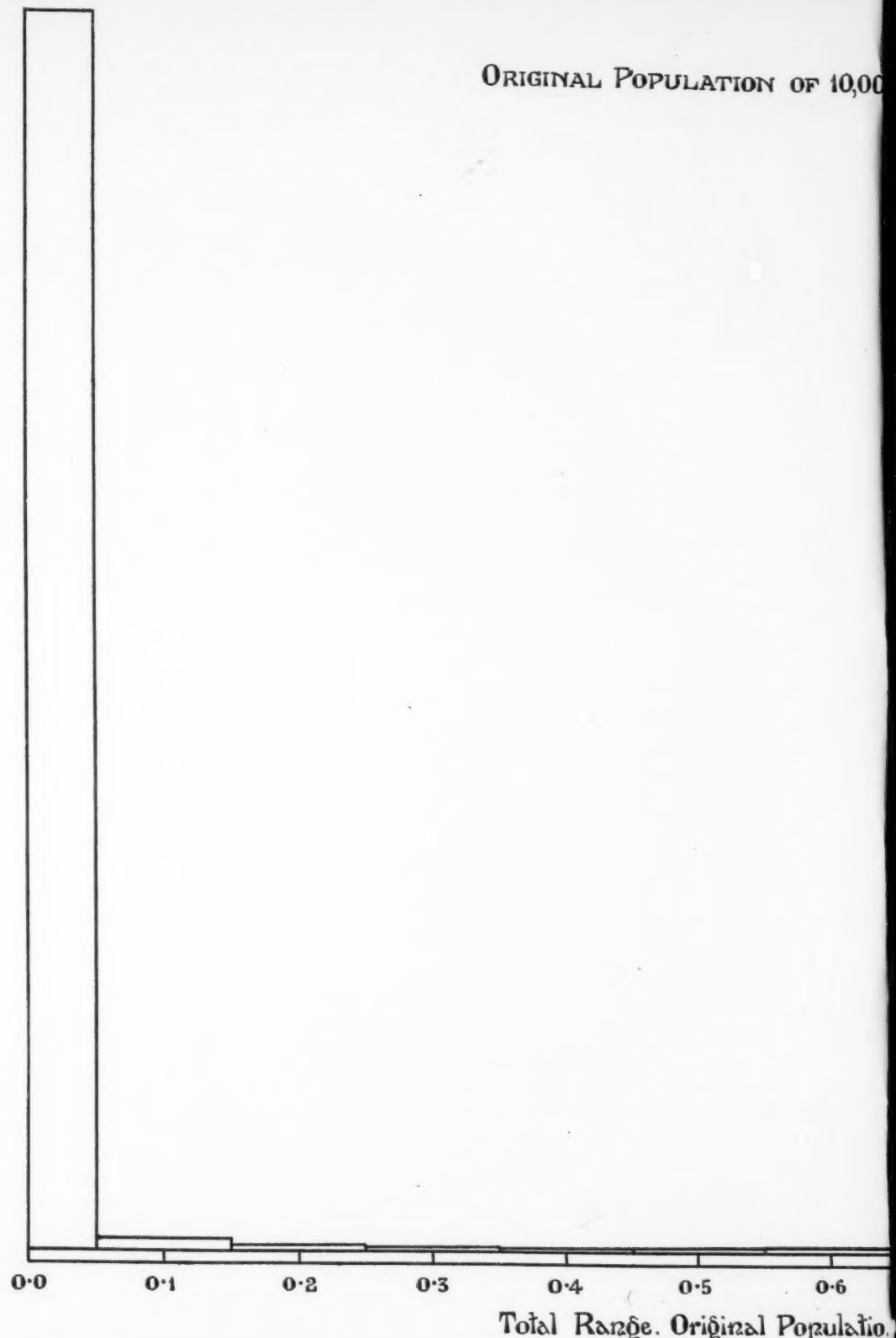
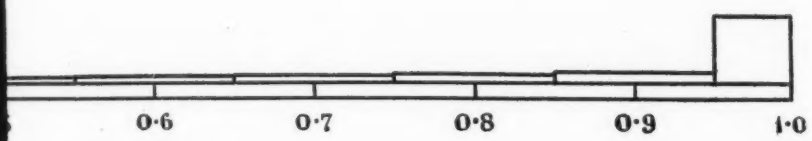

Original Population of 10,00
0·0
0·1
0·2
0·3
0·4
0·5
0·6
Total Range. Original Populatio

Fig. 1.

ATION OF 10,000.

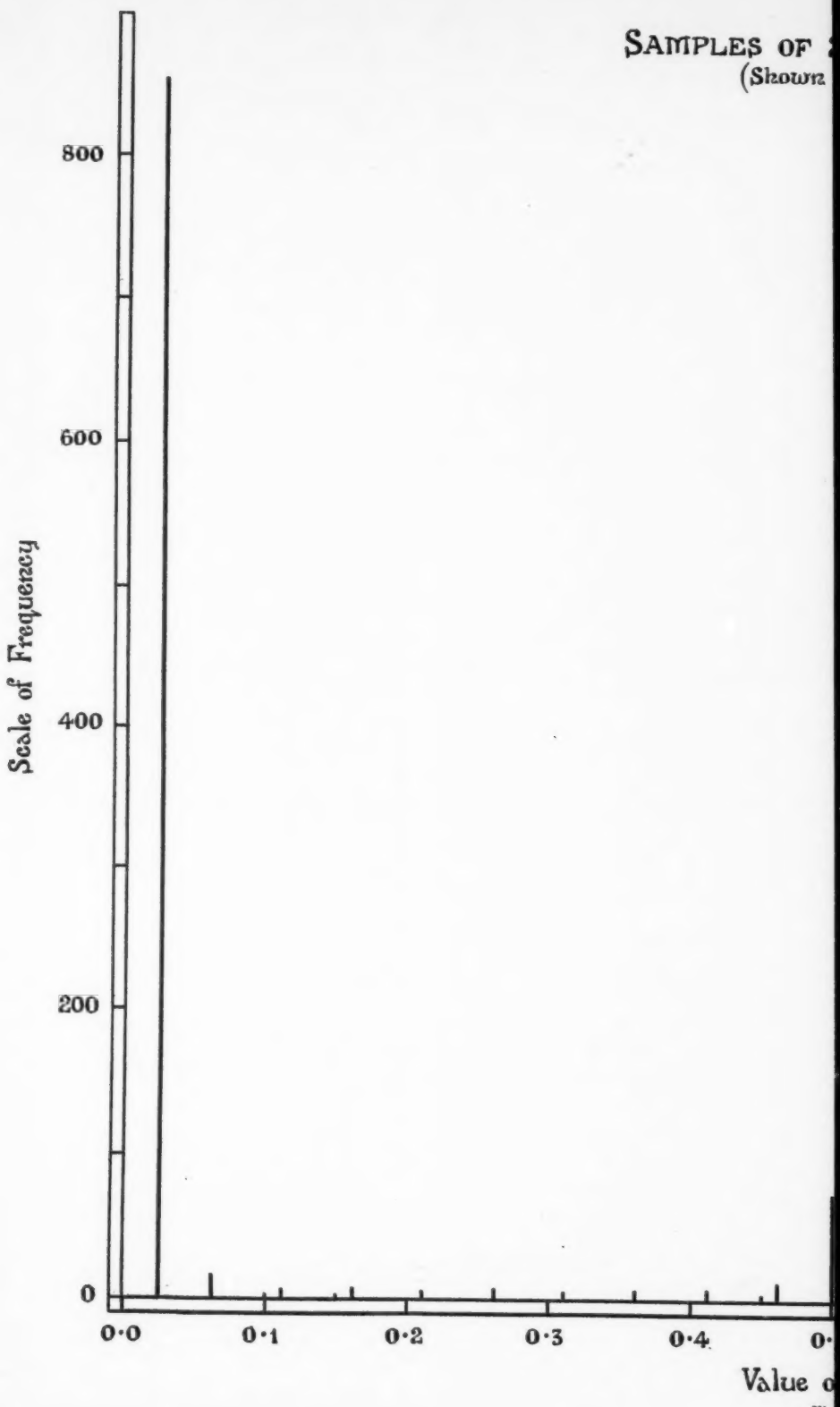
□ = 80 units

Scale of Frequency



ginal Population

SAMPLES OF
(Shown



Fig

PLES OF 2. 2000 Samples.
(Shown as 10,000)

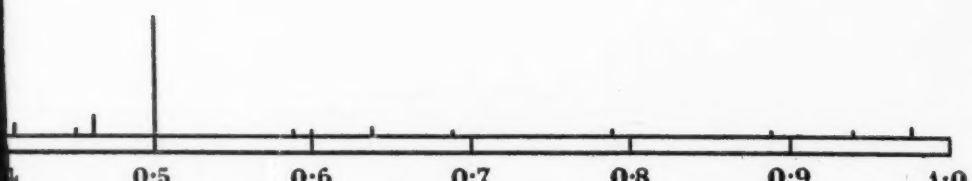


Fig. 2.

To face p. 362.

SAMPLES OF 3. (a) 1333
(Shown as 10,00

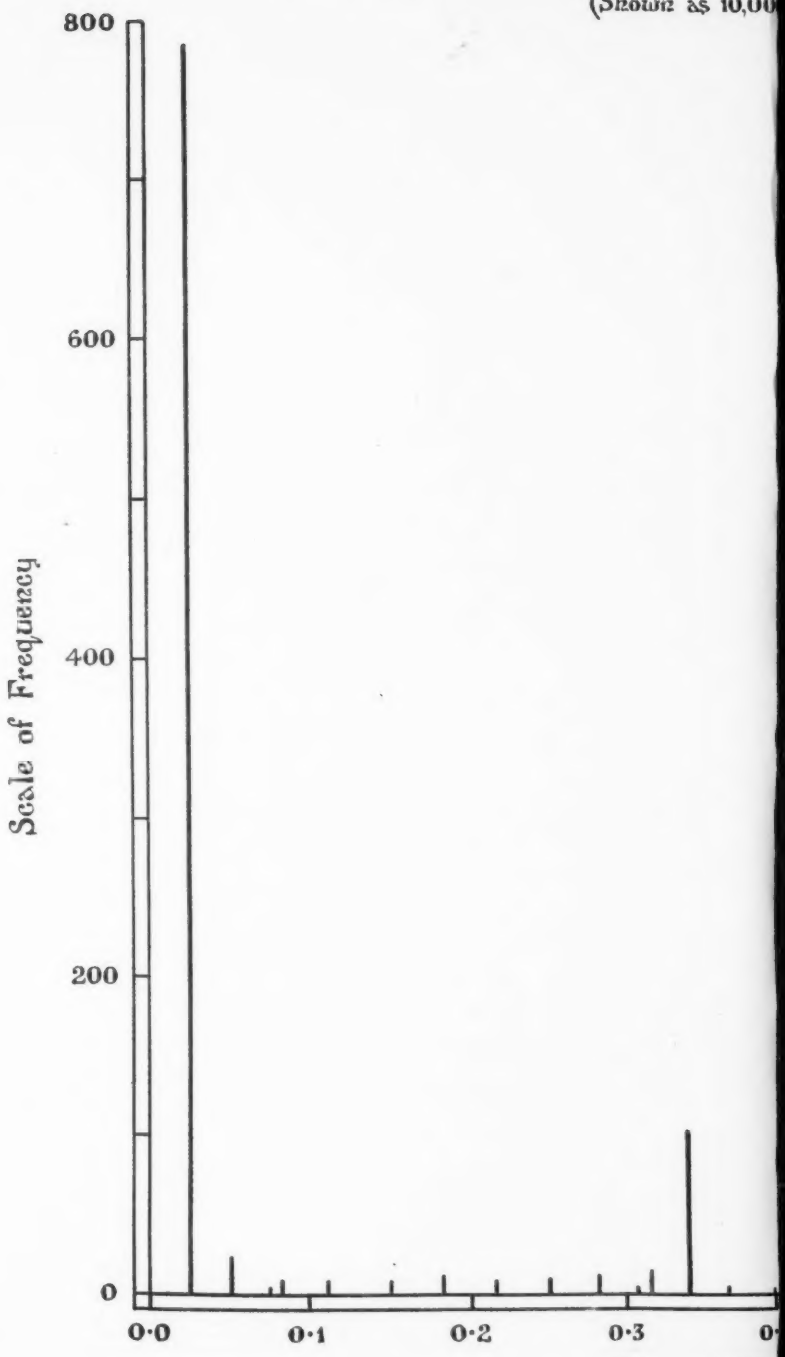
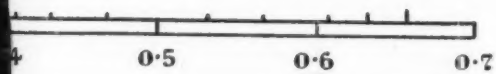


Fig. 3 (*a*).

Samples.
)



UNIV. OF TORONTO

SAMPLES OF 3. (b) 800 Sample
(Shown as 10,000)

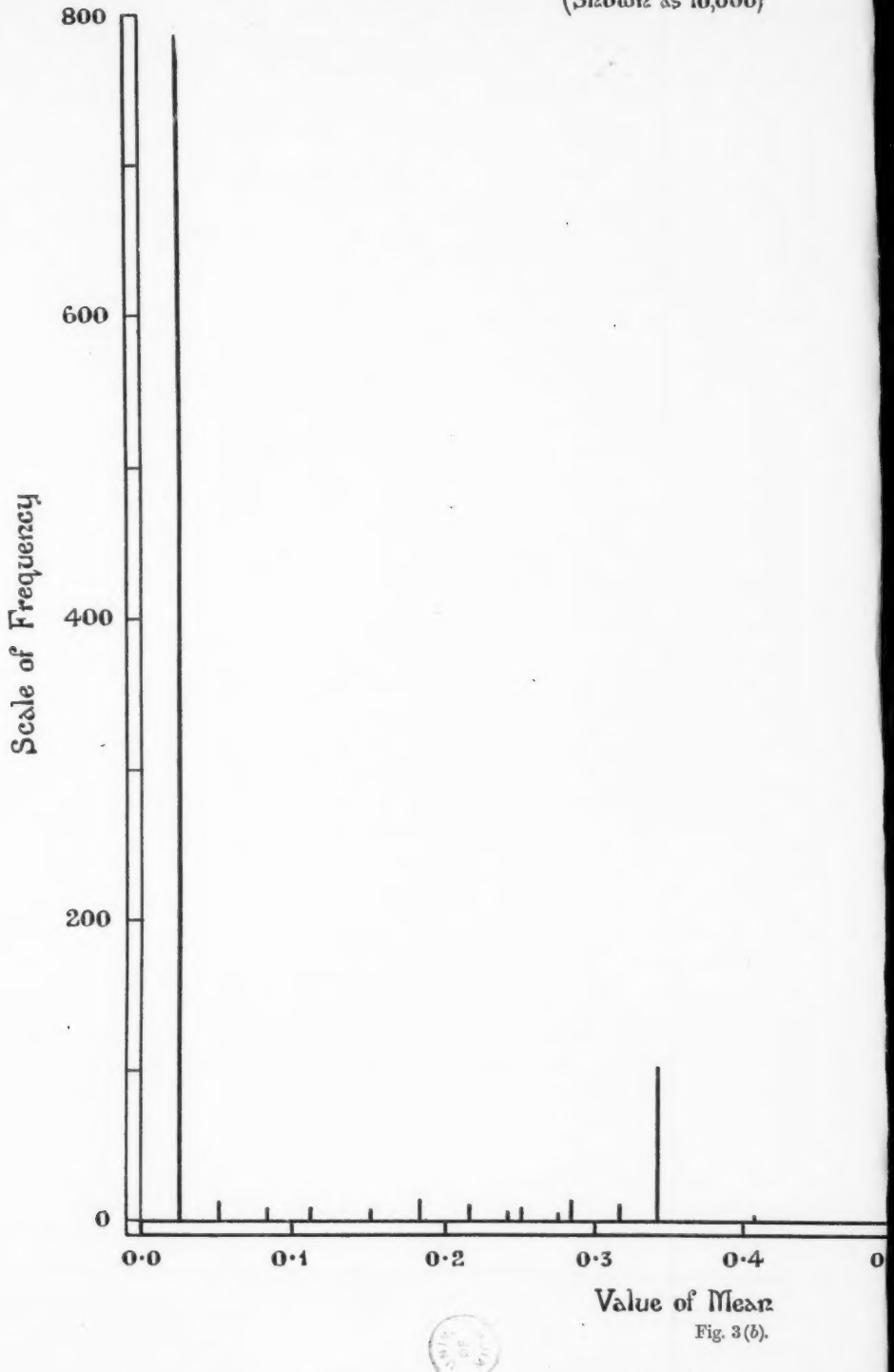
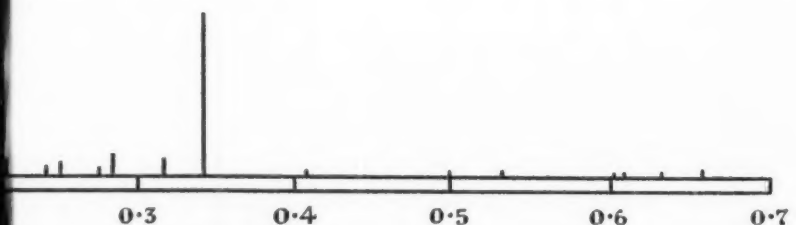


Fig. 3 (*b*).

AMPLES OF 3. (b) 800 Samples.
(Shown as 10,000)



Value of Mean

Fig. 3(*b*).

(2) The primary object of this paper is to test the efficacy of this method when the original sampled population is U-shaped. To this end a rather extreme U-shaped population was subdivided into 11 frequency groups. This number of groups is distinctly low and this probably accounts for many of the results obtained, but the labour becomes excessive if many more groups be used, whilst still further the values of the frequency in the groups, other than the first few, become relatively so small that they are useless in practice and are unrepresentable graphically.

The sampled population was as follows:

Abscissa ...	·025	·1	·2	·3	·4	·5	·6	·7	·8	·9	·975
Frequency ...	9213	84	46	35	30	28	29	31	39	64	401

with constants:

$$\text{Mean} = 0{\cdot}080{,}330, \qquad \nu_2 = 0{\cdot}0442{,}4897,$$
$$\nu_3 = -\,0{\cdot}0350{,}3332, \qquad \nu_4 = 0{\cdot}0304{,}3144,$$
$$\beta_1 = 14{\cdot}166{,}177, \qquad \beta_2 = 15{\cdot}542{,}344,$$

and
$$\sigma = 0{\cdot}210{,}3543.$$

This distribution is shown in the histogram Fig. 1 on front of folding sheet; whilst ν's are used for μ's to emphasise that, as throughout this paper, the moments are uncorrected for grouping.

The sampling was carried out entirely by the use of Tippett's *Random Sample Numbers** as this is by far the most convenient and satisfactory method, whilst the population was treated as infinite in all its categories.

The first sampling consisted of 2000 samples of 2 and gave the distribution:

Value of Mean ...	·0250	·0625	·1000	·1125	·1500	·1625	·2125	·2625	·3125	·3625	·4125
Frequency ...	1710	29	1	13	1	9	8	14	9	11	14

Value of Mean ...	·4500	·4625	·5000	·5875	·6000	·6375	·6875	·7875	·8875	·9375	·9750
Frequency ...	1	22	147	1	1	2	1	1	1	1	3

with constants:

$$\text{Mean} = 0{\cdot}079{,}250, \qquad \nu_2 = 0{\cdot}021{,}938,$$
$$\nu_3 = 0{\cdot}008{,}738, \qquad \nu_4 = 0{\cdot}004{,}514,$$
$$\beta_1 = 7{\cdot}230, \qquad \beta_2 = 9{\cdot}379,$$

and
$$\sigma = 0{\cdot}148{,}115.$$

This distribution is shown in Fig. 2 on front of folding sheet, whilst here, as elsewhere, the constants are given in terms of the unit of the sampled population.

* *Tracts for Computers* (Cambridge University Press) No. XV. A description of this method is given in the paper to which previous reference has been made.

Then these 2000 samples of 2 were combined (without fresh sampling) into 1333 samples of 3:

Value of Mean ...	·02500	·05000	·07500	·08333	·11667	·15000	·18333	·21667
Frequency ...	1050	30	1	12	9	8	14	9

Value of Mean ...	·25000	·28333	·30833	·31667	·34167	·36667	·40000	·40833
Frequency ...	10	14	1	19	134	3	2	1

Value of Mean ...	·43333	·46667	·53333	·56667	·60833	·63333	·65833
Frequency ...	2	1	1	1	1	2	8

with constants:

$$\text{Mean} = 0{\cdot}079{,}557, \qquad \nu_2 = 0{\cdot}014{,}778,$$
$$\nu_3 = 0{\cdot}003{,}963, \qquad \nu_4 = 0{\cdot}001{,}584,$$
$$\beta_1 = 4{\cdot}867, \qquad \beta_2 = 7{\cdot}254,$$

and
$$\sigma = 0{\cdot}121{,}565.$$

This distribution is shown in Fig. 3 (*a*) on back of folding sheet.

This set of samples was then combined further to give 1000 samples of 4:

Value of Mean ...	·02500	·04375	·06250	·06875	·09375	·11875	·13750	·14375
Frequency ...	729	26	1	10	9	4	2	11

Value of Mean ...	·16250	·16875	·19375	·21875	·23750	·24375	·26250
Frequency ...	1	9	9	14	1	16	128

Value of Mean ...	·28125	·30625	·31250	·33125	·35625	·38125	·40625
Frequency ...	4	4	1	2	2	2	1

Value of Mean ...	·41250	·43125	·46250	·48125	·50000	·55000	·57500
Frequency ...	1	1	1	3	6	1	1

with constants:

$$\text{Mean} = 0{\cdot}079{,}594, \quad \nu_2 = 0{\cdot}010{,}852,$$
$$\nu_3 = 0{\cdot}002{,}058, \quad \nu_4 = 0{\cdot}000{,}661,$$
$$\beta_1 = 3{\cdot}315, \quad \beta_2 = 5{\cdot}615,$$

and $$\sigma = 0{\cdot}104{,}173.$$

This distribution is shown in Fig. 4 (*a*) on p. 366.

By using 1000 more of Tippett's Numbers and combining with the previous 4000, 1000 samples of 5 were formed with the following results:

Value of Mean ...	*0·025*	*0·040*	*0·055*	*0·060*	*0·080*	*0·095*	*0·100*
Frequency ...	677	28	3	9	9	1	7

Value of Mean ...	*0·115*	*0·120*	*0·140*	*0·160*	*0·180*	*0·195*	*0·200*
Frequency ...	2	13	11	8	18	2	19

Value of Mean ...	*0·215*	*0·230*	*0·250*	*0·255*	*0·270*	*0·290*	*0·310*
Frequency ...	148	5	5	2	3	1	2

Value of Mean ...	*0·330*	*0·335*	*0·350*	*0·375*	*0·390*	*0·405*	*0·545*
Frequency ...	3	1	2	1	5	14	1

with constants:

$$\text{Mean} = 0{\cdot}079{,}320, \quad \nu_2 = 0{\cdot}008{,}629,$$
$$\nu_3 = 0{\cdot}001{,}293, \quad \nu_4 = 0{\cdot}000{,}369,$$
$$\beta_1 = 2{\cdot}603, \quad \beta_2 = 4{\cdot}953,$$

and $$\sigma = 0{\cdot}092{,}893.$$

This distribution is shown in Fig. 5 (*a*) on p. 367.

These 1000 samples of 5 were again combined with another 5000 of Tippett's Numbers to give 1000 samples of 10 as follows:

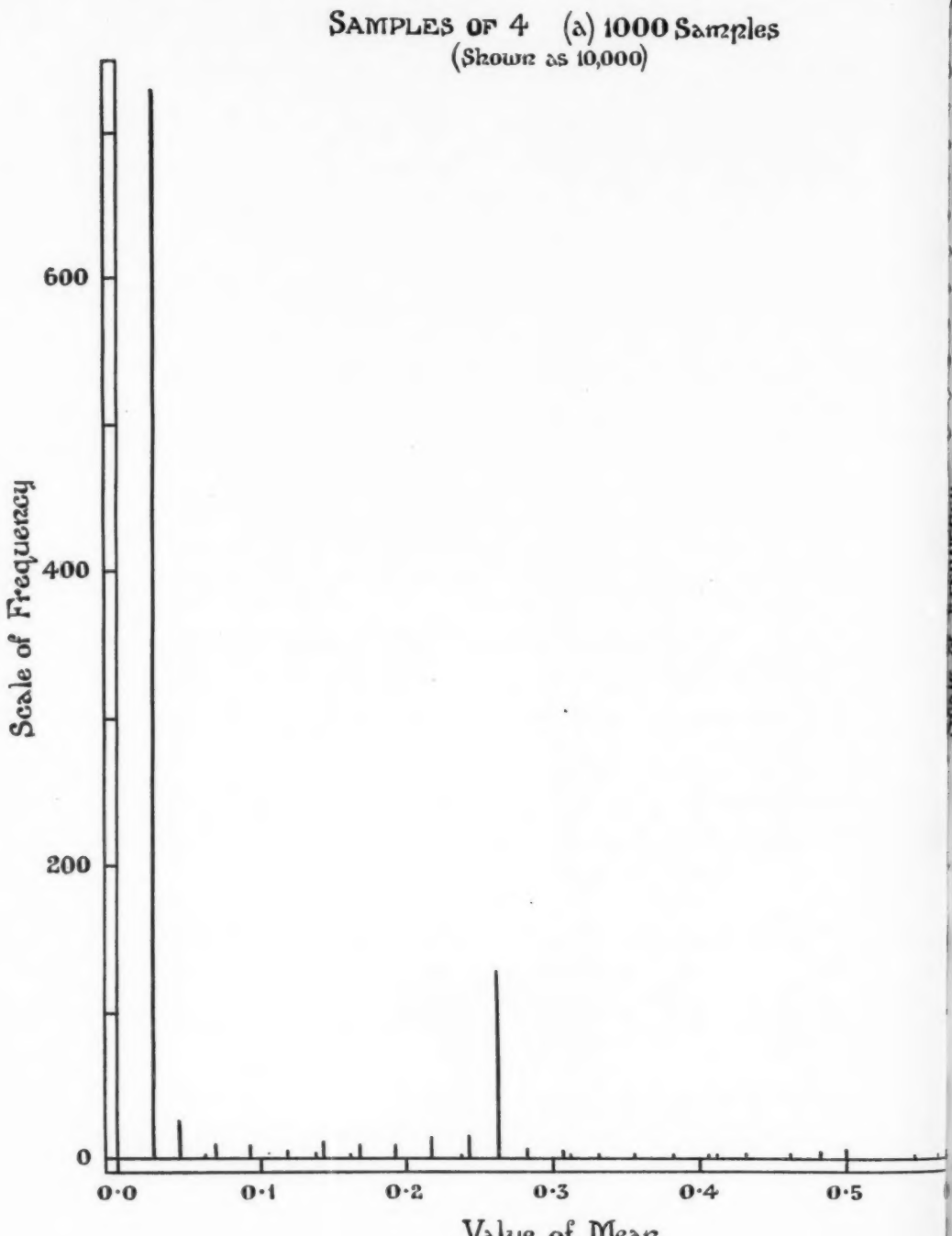

SAMPLES OF 4 (a) 1000 Samples
(Shown as 10,000)
600
400
200
0
Scale of Frequency
0·0
0·1
0·2
0·3
0·4
0·5
Value of Mean

Fig. 4 (*a*).

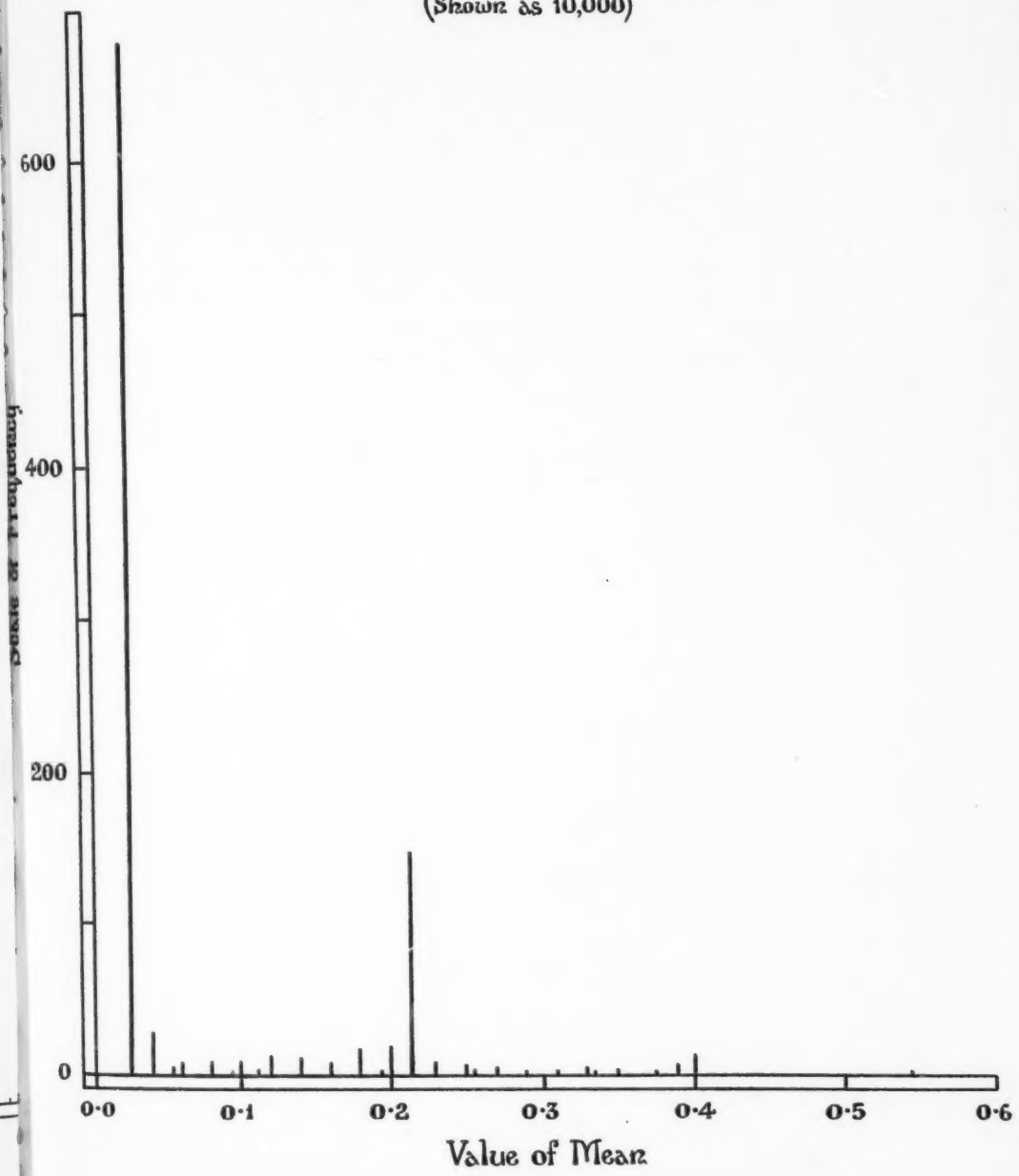

SAMPLES OF 5. (a) 1000 Samples.
(Shown as 10,000)
Frequency of Means
600
400
200
0
0·0
0·1
0·2
0·3
0·4
0·5
0·6
Value of Mean

Fig. 5 (*a*).

Value of Mean ...	·01—·03	·03—	·05—	·07—	·09—	·11—	·13—	·15—
Frequency ...	443	57	31	44	34	242	23	11

Value of Mean ...	·17—	·19—	·21—	·23—	·25—	·27—	·29—	·31—·33
Frequency ...	17	26	50	8	5	4	1	4

with constants*:

$$\text{Mean} = 0{\cdot}079{,}120, \quad \nu_2 = 11{\cdot}454{,}064,$$
$$\nu_3 = 37{\cdot}234{,}022, \quad \nu_4 = 419{\cdot}360{,}134,$$
$$\beta_1 = 0{\cdot}9226, \quad \beta_2 = 3{\cdot}196,$$

and
$$\sigma = 0{\cdot}067{,}6877.$$

This distribution is shown in Fig. 6 (*a*) below.

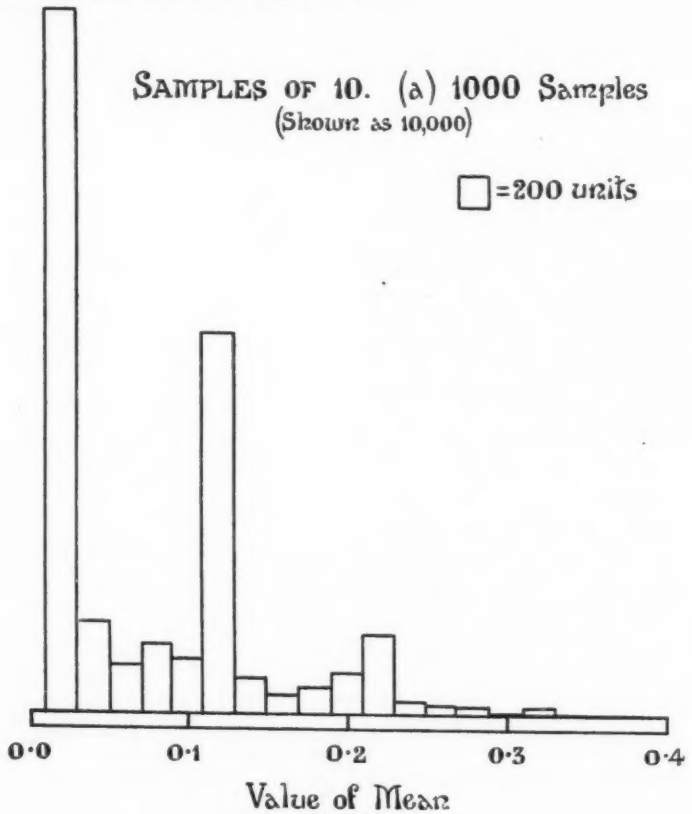


Fig. 6 (*a*).

* To save the long array of zeros the values of ν_2, ν_3 and ν_4 are here given in working units.

These 1000 samples of 10 were then combined to form 400 samples of 25:

Value of Mean ...	0·02—0·03	0·03—	0·04—	0·05—	0·06—	0·07—	0·08—	0·09—	0·10—
Frequency ...	69	8	27	11	90	22	19	21	46

Value of Mean ...	0·11—	0·12—	0·13—	0·14—	0·15—	0·16—	0·17—	0·18—	0·19—0·20
Frequency ...	17	17	26	7	6	4	5	4	1

with constants*:

$$\text{Mean} = 0{\cdot}079{,}425, \qquad \nu_2 = 16{\cdot}201{,}694,$$
$$\nu_3 = 30{\cdot}043{,}111, \qquad \nu_4 = 671{\cdot}018{,}480,$$
$$\beta_1 = 0{\cdot}2122, \qquad \beta_2 = 2{\cdot}556,$$

and
$$\sigma = 0{\cdot}040{,}2513.$$

This distribution is shown in Fig. 7 (*a*) on p. 370.

Finally the 400 samples of 25 were combined to give 200 samples of 50:

Value of Mean ...	0·02—0·03	0·03—	0·04—	0·05—	0·06—	0·07—	0·08—	0·09—
Frequency ...	10	8	15	18	27	20	34	23

Value of Mean ...	0·10—	0·11—	0·12—	0·13—	0·14—	0·15—	0·16—0·17
Frequency ...	18	8	11	5	2	0	1

with constants:

$$\text{Mean} = 0{\cdot}080{,}100, \qquad \nu_2 = 7{\cdot}869{,}900,$$
$$\nu_3 = 1{\cdot}926{,}402, \qquad \nu_4 = 175{\cdot}270{,}722,$$
$$\beta_1 = 0{\cdot}007{,}614, \qquad \beta_2 = 2{\cdot}830,$$

and
$$\sigma = 0{\cdot}028{,}0533.$$

This distribution is shown in Fig. 8 (*a*) on p. 370.

Now using the formulae ${}_1B_1 = \dfrac{\bar{\beta}_1}{N}$ and ${}_1B_2 - 3 = \dfrac{\bar{\beta}_2 - 3}{N}$ we can calculate im-

* See footnote on p. 368.

mediately the theoretical values of the betas of these distributions of samples of 2, 3, 4, 5, 10, 25 and of 50, and we have, since $\beta_1 = 14{\cdot}166{,}177$, $\beta_2 = 15{\cdot}542{,}344$,

N	${}_1B_1$	${}_1B_2$
2	7·083	9·271
3	4·722	7·181
4	3·515	6·136
5	2·833	5·508
10	1·417	4·254
25	0·567	3·502
50	0·280	3·251

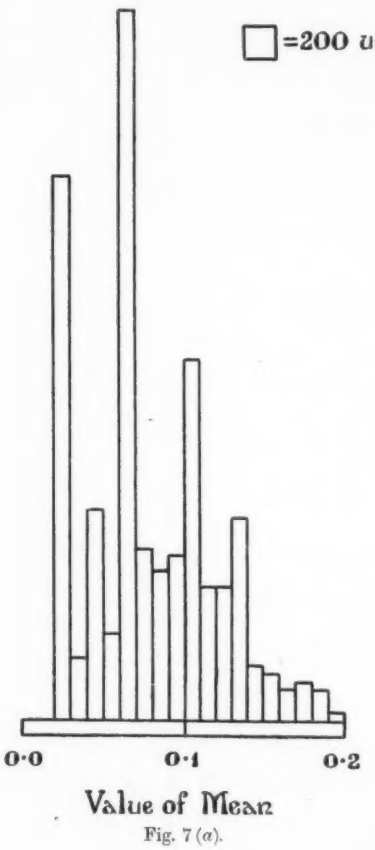


Fig. 7 (*a*).

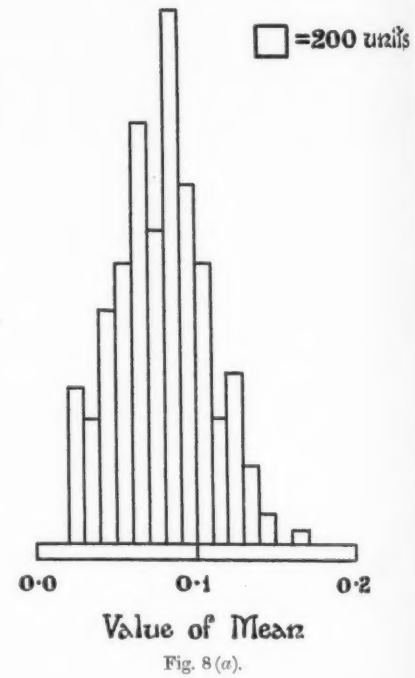


Fig. 8 (*a*).

which can be compared with the β_1 and β_2 of the distributions of means just given.

Fig. 9 (*a*) below shows this comparison. The points representing the theoretic values of the pairs of betas are the centres of the ellipses, whilst the large dots represent the pairs of values of the betas obtained by actual sampling.

To obtain some criterion as to the significance of the deviations of the values of the actual betas from their theoretic values the probability ellipses were drawn

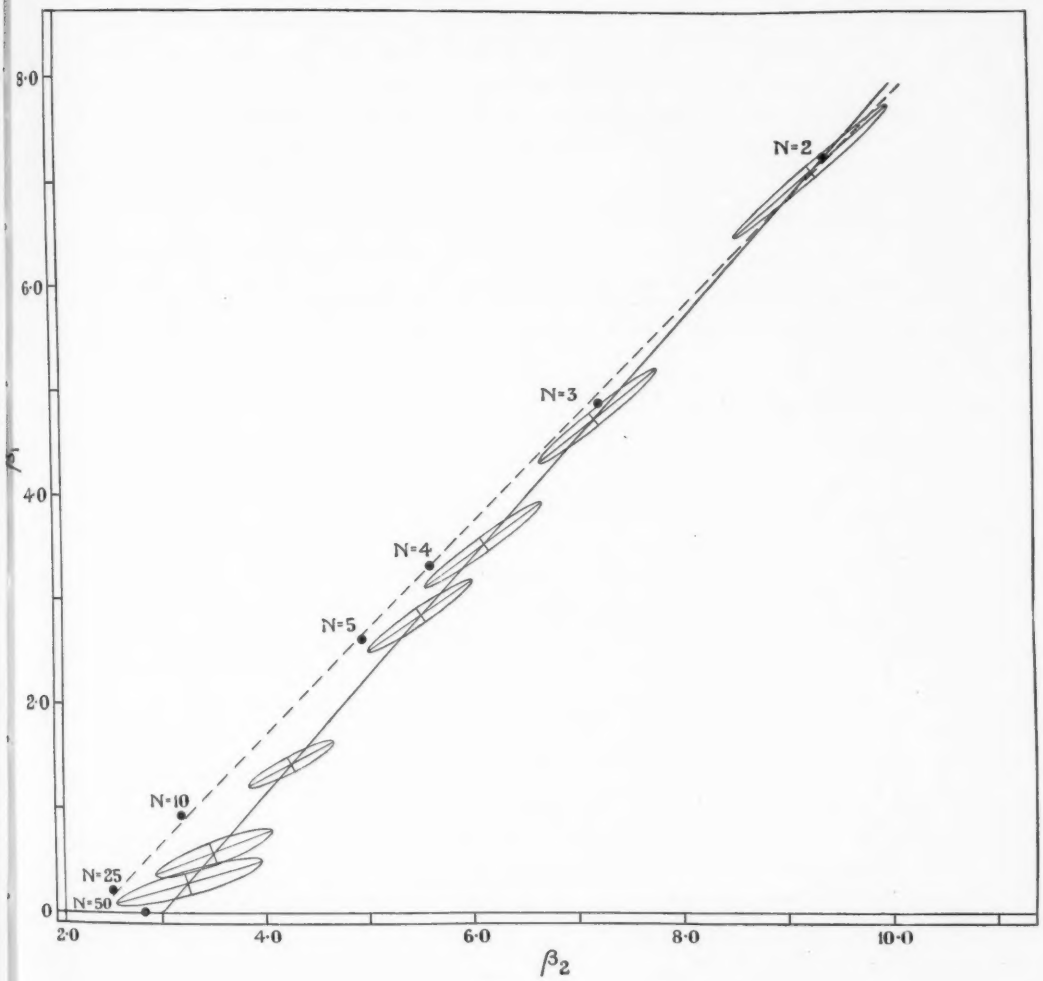


Fig. 9 (*a*).

as shown, but it must be emphasised that they have only partial validity, they give merely a rough indication as to whether or no an actual pair of betas is a likely approximation to the corresponding theoretical pair of values. The axes of the ellipses were calculated, by assuming that the higher betas corresponding to a theoretical pair ${}_1\beta_1$, ${}_1\beta_2$ are given by the formulae

$$\beta_n\,(\text{even}) = (n+1)\,\{\tfrac{1}{2}\beta_{n-1} + (1+\tfrac{1}{2}\alpha)\,\beta_{n-2}\}/[1-\tfrac{1}{2}\,(n-1)\,\alpha],$$

$$\beta_n\,(\text{odd}) = (n+1)\,\{\tfrac{1}{2}\beta_1\beta_{n-1} + (1+\tfrac{1}{2}\alpha)\,\beta_{n-2}\}/[1-\tfrac{1}{2}\,(n-1)\,\alpha],$$

where
$$\alpha = \frac{2\beta_2 - 3\beta_1 - 6}{\beta_2 + 3},$$

and then by using the formula for *normal* correlation,

$$(1{\cdot}1774)^2 = \frac{1}{1-R^2_{\beta_1\beta_2}}\left(\frac{\beta_1{}^2}{\Sigma^2_{\beta_1}} + \frac{\beta_2{}^2}{\Sigma^2_{\beta_2}} - \frac{2R_{\beta_1\beta_2}\cdot\beta_1\beta_2}{\Sigma_{\beta_1}\,\Sigma_{\beta_2}}\right)$$

of the *Tables for Statisticians*, p. lxiv, this being the contour inside which 50 °/₀ of the ${}_1\beta_1$, ${}_1\beta_2$ values should fall in repeated random sampling.

The lengths and inclinations to the axis of β_2 of the axes of these ellipses are:

Samples of 2	2000 Samples	Semi-Major axis	= 0·98
		Semi-Minor axis	= 0·06
		Inclination of Major axis	= 40° 40′
Samples of 3	1333 Samples	Semi-Major axis	= 0·77
		Semi-Minor axis	= 0·07
		Inclination of Major axis	= 38° 21′
Samples of 4	1000 Samples	Semi-Major axis	= 0·70
		Semi-Minor axis	= 0·07
		Inclination of Major axis	= 36° 0′
Samples of 5	1000 Samples	Semi-Major axis	= 0·61
		Semi-Minor axis	= 0·07
		Inclination of Major axis	= 34° 48′
Samples of 10	1000 Samples	Semi-Major axis	= 0·46
		Semi-Minor axis	= 0·07
		Inclination of Major axis	= 28° 16′
Samples of 25	400 Samples	Semi-Major axis	= 0·60
		Semi-Minor axis	= 0·11
		Inclination of Major axis	= 20° 19′
Samples of 50	200 Samples	Semi-Major axis	= 0·74
		Semi-Minor axis	= 0·13
		Inclination of Major axis	= 16° 10′

We can now give consideration to the primary object of this work. From Fig. 9(*a*) it is clear that, in the majority of the cases the betas of the actual distributions of means of samples are moderately close to their theoretical values, but an examination of Figs. 2 to 7 (*a*) shows immediately that, until $N = 50$, these distributions are quite unrepresentable by any Pearson Curve.

We have here evidence that, though the betas and the use of a Pearson Curve may give a good representation of a distribution of means of small samples from an ordinary skew population, yet when that sampled population is *U*-shaped the method fails until the size of the sample has become at least moderately large. It will be seen that the second series of distributions of means in samples, dealt with later in this paper, tends to confirm this view.

Returning again to Fig. 1 it is easily seen that the points representing actual pairs of betas are practically collinear from $N = 2$ to $N = 25$, the line being shown dotted in the figure. This dotted line is a representation of the best fitting straight line to these points and its equation is approximately

$$\beta_2 = 0{\cdot}986\beta_1 + 2{\cdot}352,$$

and the figure shows the manner in which it deviates distinctly from the true line of betas, i.e. ${}_1B_2 = \text{const.} \times {}_1B_1 + 3$ which is, in this case, $\beta_2 = 0{\cdot}885\beta_1 + 3$, the ordinary notation for the betas being now used as there is no question of confusion.

This peculiarity of the betas of the actual distributions of samples of N must be considered in connection with the fact that each set of samples of N was obtained from the previous set by the regrouping of the elements of the samples together with additions to their number when necessary. Thus the existence of such a line of collinearity would seem to indicate that there is some relation between one set of samples and the preceding set from which it was derived. That is, although when sets of samples of N are required for different values of N it is much more expeditious to obtain any set by regrouping the elements of the samples of the previous set corresponding to a smaller value of N, such a method would appear to be unsatisfactory from a sampling point of view because there is a loss of randomness in sets of samples so formed.

(3) Thus it is indicated that, in order to obtain truly random sets of samples of N by the use of Tippett's Random Sampling Numbers it is necessary to make a fresh allocation of the numbers for each set of samples. With a view to obtaining further evidence upon this point and also on our primary object, the representation of the distributions formed by such sets of samples, it was decided to obtain another set of samples of 800 each corresponding to $N = 3$, 4, 5, 10, 25 and 50, each set of samples being rendered independent of the preceding set by a re-allocation each time of the numbers in Tippett's Table.

The distributions of the means of the sets of samples of 800, each obtained from Tippett's Table in this manner, are as follows:

800 *Samples of* 3.

Value of Mean ...	*0·0250*	*0·0500*	*0·0833*	*0·1167*	*0·1500*	*0·1833*	*0·2167*
Frequency ...	628	13	4	6	4	7	8

Value of Mean ...	*0·2500*	*0·2833*	*0·3083*	*0·3167*	*0·3417*	*0·3667*	*0·4417*
Frequency ...	7	8	1	14	90	2	1

Value of Mean ...	*0·4667*	*0·4750*	*0·5000*	*0·5667*	*0·6000*	*0·6083*	*0·6583*
Frequency ...	1	1	1	1	1	1	1

with constants:

$$\text{Mean} = 0{\cdot}084{,}585, \qquad \nu_2 = 0{\cdot}016{,}434,$$
$$\nu_3 = 0{\cdot}004{,}337, \qquad \nu_4 = 0{\cdot}001{,}701,$$
$$\beta_1 = 4{\cdot}237, \qquad \beta_2 = 6{\cdot}297,$$

and
$$\sigma = 0{\cdot}128{,}195.$$

This distribution is shown in Fig. 3 (*b*) on the back of folding sheet, p. 363.

800 *Samples of* 4.

Value of Mean ...	*0·02500*	*0·04375*	*0·06250*	*0·06875*	*0·09375*	*0·11875*	*0·14375*	*0·16875*
Frequency ...	587	23	1	10	7	10	6	6

Value of Mean ...	*0·19375*	*0·21875*	*0·24375*	*0·26250*	*0·28125*	*0·33125*	*0·35625*	*0·38125*
Frequency ...	12	10	11	103	1	2	1	1

Value of Mean ...	*0·38750*	*0·40625*	*0·43750*	*0·45625*	*0·46250*	*0·48125*	*0·50000*
Frequency ...	1	1	1	1	1	1	3

with constants:

$$\text{Mean} = 0{\cdot}075{,}242, \qquad \nu_2 = 0{\cdot}009{,}490,$$
$$\nu_3 = 0{\cdot}001{,}661, \qquad \nu_4 = 0{\cdot}000{,}474,$$
$$\beta_1 = 3{\cdot}227, \qquad \beta_2 = 5{\cdot}260,$$

and
$$\sigma = 0{\cdot}097{,}417.$$

This distribution is shown in Fig. 4 (*b*) on p. 376.

800 *Samples of* 5.

Value of Mean ...	*0·025*	*0·040*	*0·060*	*0·080*	*0·100*	*0·120*	*0·140*	*0·160*
Frequency ...	531	30	13	9	5	7	10	14

Value of Mean ...	*0·180*	*0·195*	*0·200*	*0·215*	*0·230*	*0·250*	*0·275*	*0·295*
Frequency ...	18	2	23	109	5	2	1	1

Value of Mean ...	*0·310*	*0·330*	*0·335*	*0·365*	*0·375*	*0·405*	*0·580*
Frequency ...	2	3	1	1	1	11	1

with constants:

$$\text{Mean} = 0{\cdot}078{,}200, \qquad \nu_2 = 0{\cdot}008{,}141,$$
$$\nu_3 = 0{\cdot}001{,}209, \qquad \nu_4 = 0{\cdot}000{,}351,$$
$$\beta_1 = 2{\cdot}707, \qquad \beta_2 = 5{\cdot}302,$$

and
$$\sigma = 0{\cdot}090{,}227.$$

This distribution is shown in Fig. 5 (*b*) on p. 377.

800 *Samples of* 10.

Value of Mean ...	*0·0000—0·0320*	*0·0320—*	*0·0620—*	*0·0920—*	*0·1220—*	*0·1520—*	*0·1820—*
Frequency ...	307	76	31	229	37	25	24

Value of Mean ...	*0·2120—*	*0·2420—*	*0·2720—*	*0·3020—*	*0·3320—*	*0·3620—0·3920*
Frequency ...	48	4	3	14	0	2

with constants:

$$\text{Mean} = 0{\cdot}087{,}600, \qquad \nu_2 = 5{\cdot}588{,}596,$$
$$\nu_3 = 16{\cdot}679{,}491, \qquad \nu_4 = 136{\cdot}648{,}581,$$
$$\beta_1 = 1{\cdot}594, \qquad \beta_2 = 4{\cdot}375,$$

and
$$\sigma = 0{\cdot}070{,}9206.$$

This distribution is shown in Fig. 6 (*b*) on p. 378.

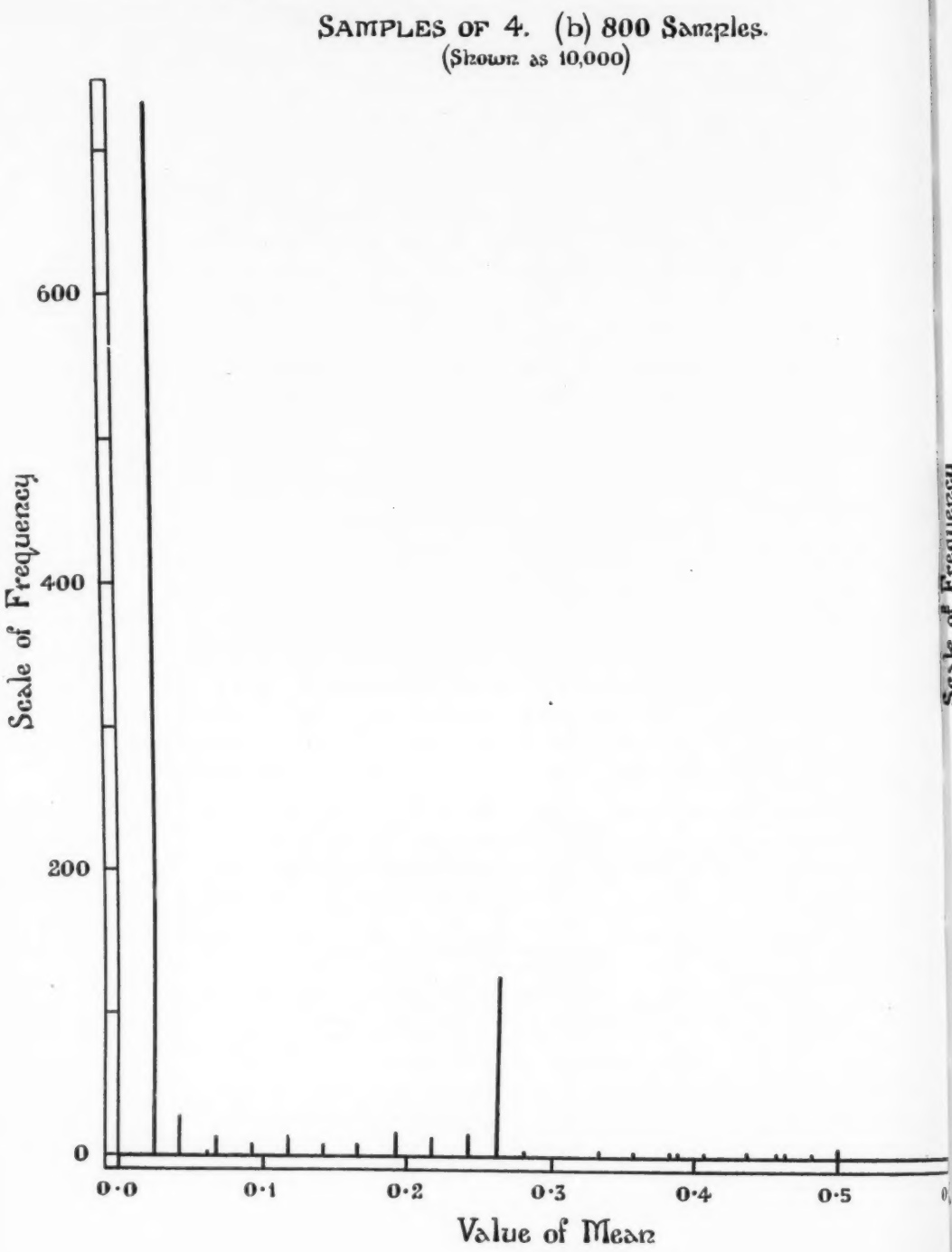

SAMPLES OF 4. (b) 800 Samples.
(Shown as 10,000)
600
400
200
0
Scale of Frequency
0·0
0·1
0·2
0·3
0·4
0·5
Value of Mean

Fig. 4 (*b*).

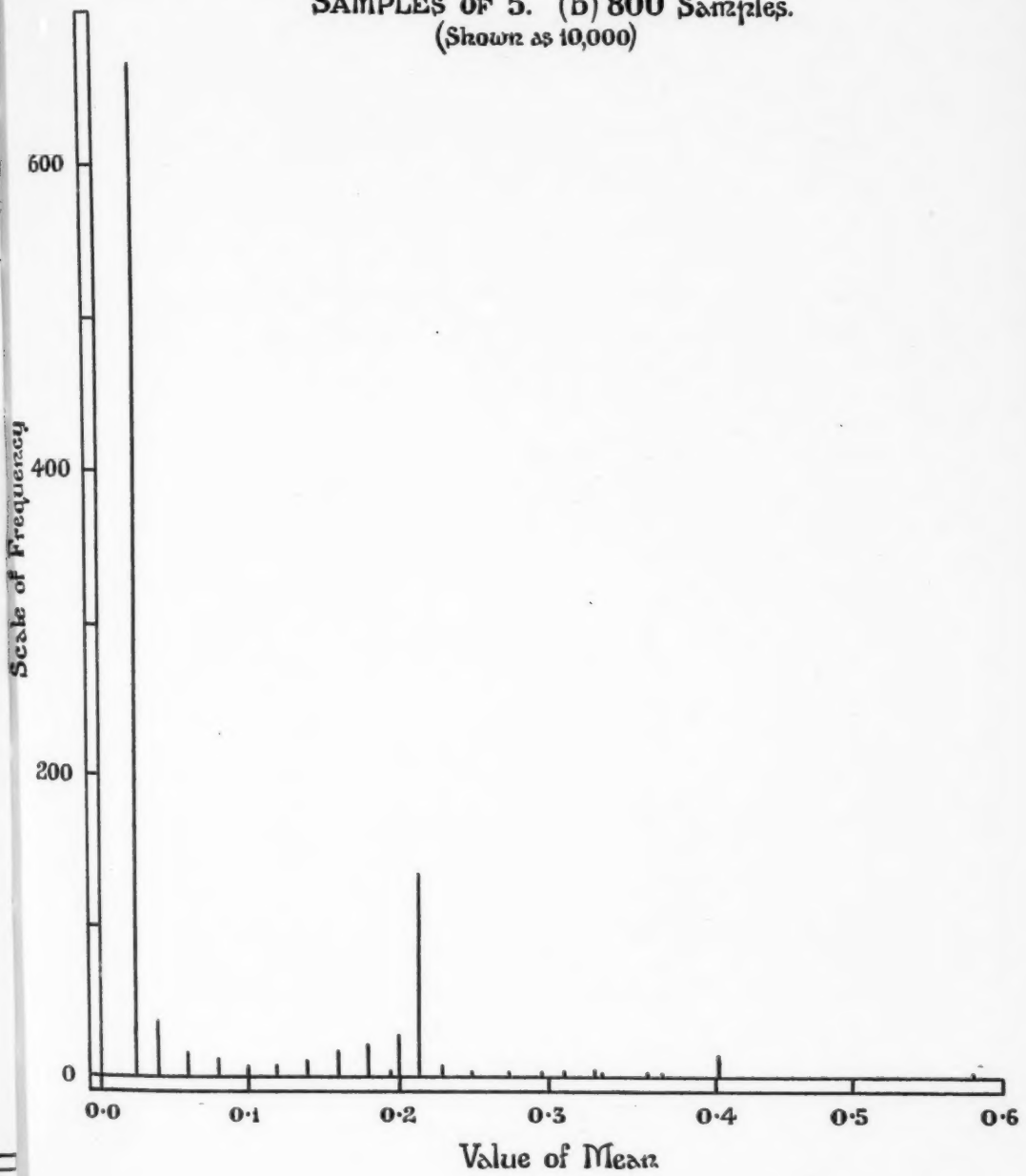

SAMPLES OF 5. (b) 800 Samples.
(Shown as 10,000)
600
400
200
0
Scale of Frequency
0·0
0·1
0·2
0·3
0·4
0·5
0·6
Value of Mean

Fig. 5 (*b*).

800 *Samples of* 25.

Value of Mean ...	*0·02—0·03*	*0·03—*	*0·04—*	*0·05—*	*0·06—*	*0·07—*	*0·08—*	*0·09—*	*0·10—*
Frequency ...	138	46	28	54	169	44	44	51	87

Value of Mean ...	*0·11—*	*0·12—*	*0·13—*	*0·14—*	*0·15—*	*0·16—*	*0·17—*	*0·18—*	*0·19—*
Frequency ...	28	20	38	15	5	10	14	3	1

Value of Mean ...	*0·20—*	*0·21—*	*0·22—*	*0·23—*	*0·24—*	*0·25—*	*0·26—0·27*
Frequency ...	0	3	0	0	1	0	1

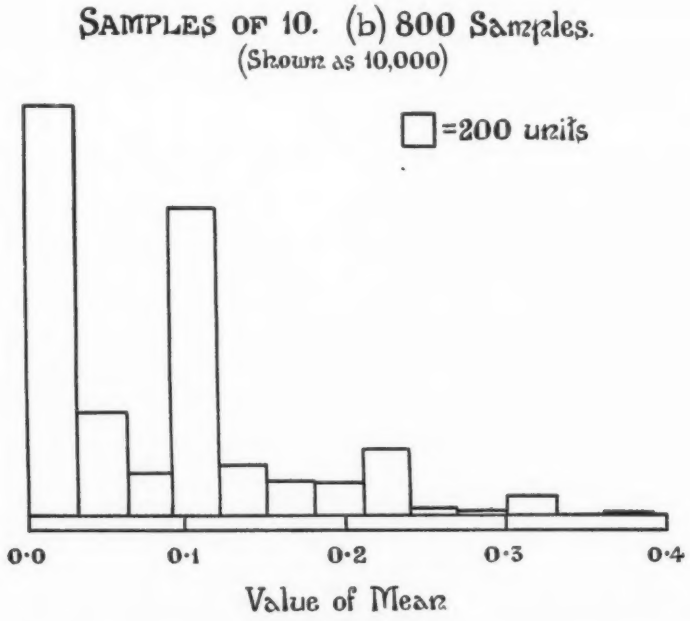


Fig. 6 (*b*).

with constants:

$$\text{Mean} = 0{\cdot}066{,}375, \qquad \nu_2 = 16{\cdot}456{,}094,$$

$$\nu_3 = 53{\cdot}796{,}761, \qquad \nu_4 = 1{,}011{\cdot}049{,}689,$$

$$\beta_1 = 0{\cdot}649, \qquad \beta_2 = 3{\cdot}733,$$

and $$\sigma = 0{\cdot}040{,}5661.$$

This distribution is shown in Fig. 7 (*b*) below.

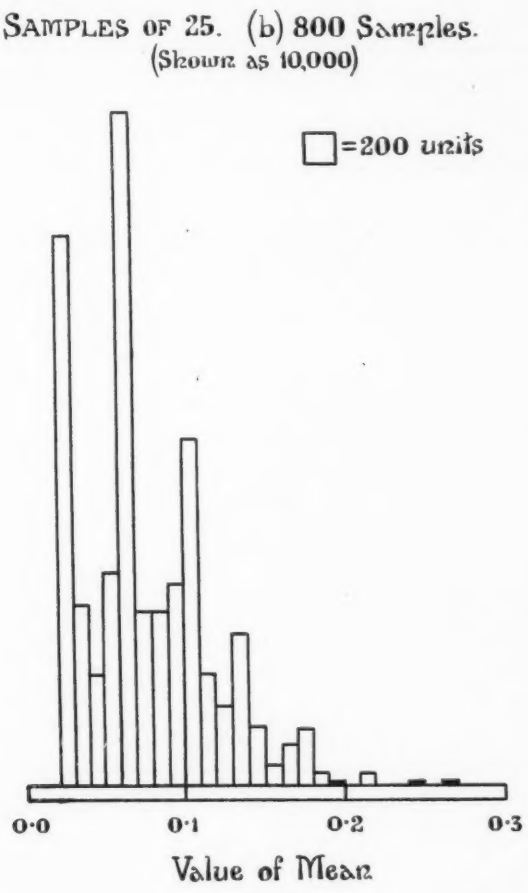


Fig. 7 (*b*).

800 *Samples of* 50.

Value of Mean ...	*0·02225—0·03225*	*0·03225—*	*0·04225—*	*0·05225—*	*0·06225—*
Frequency ...	35	31	106	71	120

Value of Mean ...	*0·07225—*	*0·08225—*	*0·09225—*	*0·10225—*	*0·11225—*	*0·12225—*
Frequency ...	122	75	80	52	45	24

Value of Mean ...	*0·13225—*	*0·14225—*	*0·15225—*	*0·16225—*	*0·17225—0·18225*
Frequency ...	15	11	8	3	2

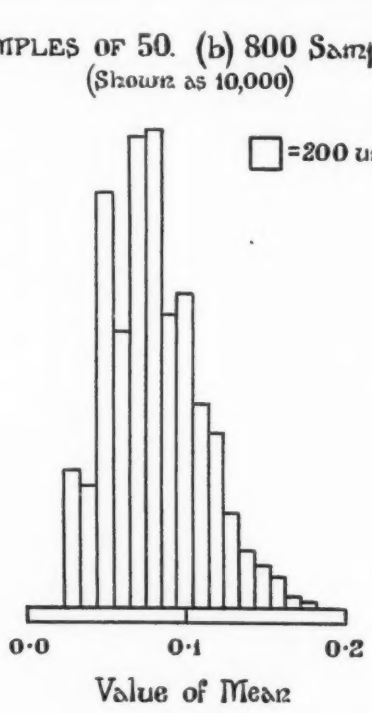


Fig 8 (*b*).

with constants:

$$\text{Mean} = 0{\cdot}072{,}375, \quad \nu_2 = 8{\cdot}669{,}844,$$
$$\nu_3 = 13{\cdot}805{,}005, \quad \nu_4 = 230{\cdot}681{,}732,$$
$$\beta_1 = 0{\cdot}292, \quad \beta_2 = 3{\cdot}069,$$

and

$$\sigma = 0{\cdot}029{,}440.$$

This distribution is shown in Fig. 8 (*b*) on p. 380.

Fig. 9 (*b*) below shows the values of the betas of these actual distributions, the corresponding theoretical values, and the probability ellipses which give some

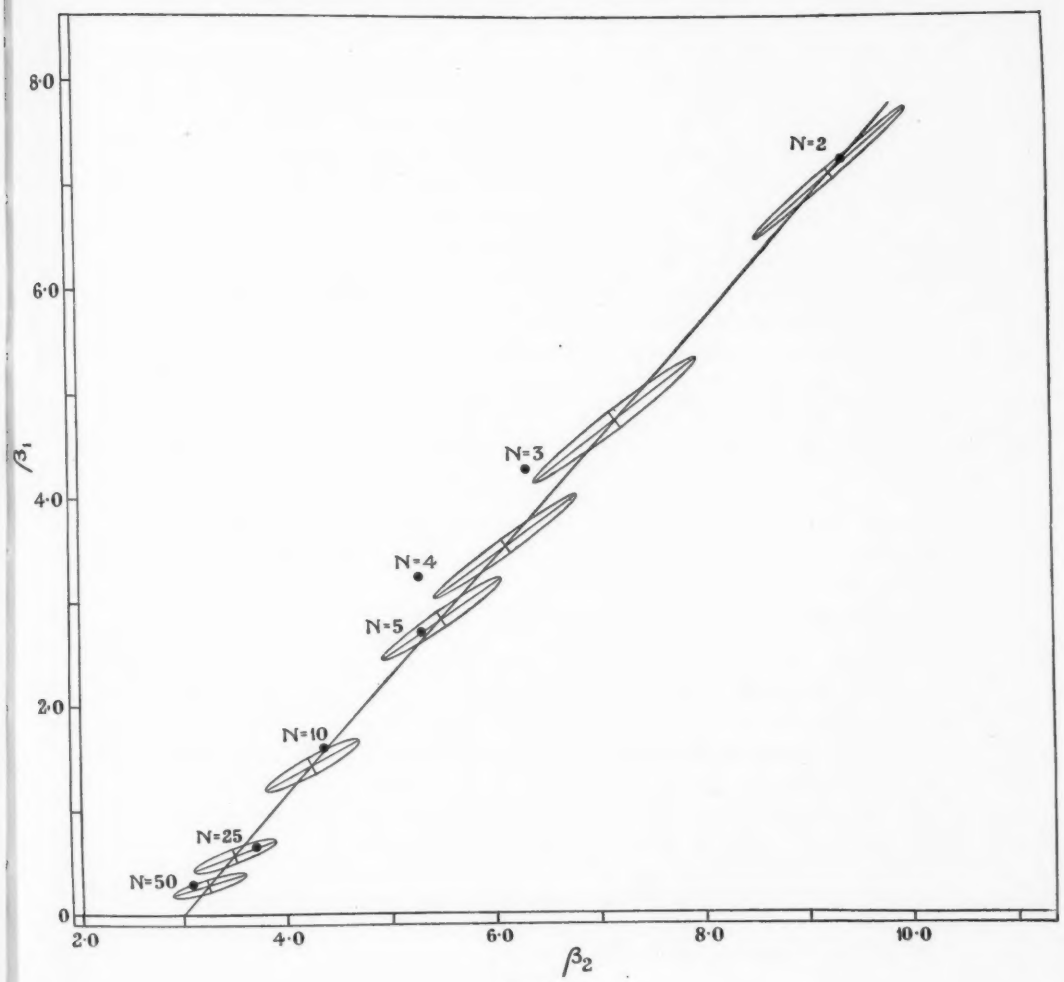


Fig. 9 (*b*).

indication of the size of the deviations from theory. As the sizes of the samples are the same as before the theoretical values of the betas are the same, but as the number of samples in each set is now 800, except for $N=2$, the probability ellipses are different.

Their axes are:

Samples of 3	Semi-Major axis	= 0·99
	Semi-Minor axis	= 0·09
Samples of 4	Semi-Major axis	= 0·85
	Semi-Minor axis	= 0·08
Samples of 5	Semi-Major axis	= 0·69
	Semi-Minor axis	= 0·08
Samples of 10	Semi-Major axis	= 0·51
	Semi-Minor axis	= 0·08
Samples of 25	Semi-Major axis	= 0·43
	Semi-Minor axis	= 0·07
Samples of 50	Semi-Major axis	= 0·37
	Semi-Minor axis	= 0·06

The inclinations of the axes to the axis of β_2 are the same as before.

Examining first Figs. 3 to 7 (*b*) for the distribution of the means in these sets of samples, we are confirmed in our view that the values of the betas and the use of a Pearson Curve will fail to give representations of these distributions when N is less than 50, which was the first point on which we wished to be clear.

Now considering the randomness or otherwise of the sampling it is clear from the figure that the points representing the betas of these actual distributions of means are no longer collinear, thus confirming the view that the previous close collinearity of these points was connected with the fact that each set of samples was derived from the preceding set. There still is a tendency for the points representing the actual betas to be above the line of theoretical betas but they are not collinear nor are they uniformly above the theoretical line, whilst, as one would expect if the sampling is good, the actual and theoretical values become closer as N increases. Further considering the probability ellipses, we note that $3\frac{1}{2}$ points lie inside and $3\frac{1}{2}$ points lie outside the probability-ellipse areas, a reasonable approximation to 50 %.

Let us consider the distribution of means either as represented in the diagrams or in the tables a little more in detail. They suggest a tendency to assume multiple U-forms with marked maxima at the connecting points. For example in the case of $N=2$, the form appears to be that of two U-curves with junction at a value 0·5 of the mean. A certain amount of irregularity is of course due to the fact that owing to the limited number of population groups the mean n-samples can only assume certain discrete values which occur at unequal intervals, but they cannot account for the main maxima.

As pointed out in an Editorial Note at the end of Mr Philip Hall's paper on the Distribution of Means of Samples from a Rectangular Population*, it is probable that the distribution of means in samples from a continuous limited range population will consist of a number of small arcs connected at the points $\alpha + mk/N$ $(m = 1, 2 \ldots N-1)$ where k is the range in the sampled population and α the distance of the commencement of the curve from the origin. It is easy to see that in sampling from a continuous U-shaped curve such arcs are likely to give maxima at their points of junction, and this effect is likely to persist when the population has been divided into a rather limited number of groups.

In the present case $\alpha = \cdot025$ and $k = \cdot950$, from which the following results are obtained.

(a) *Expected points of junction of arcs for a continuous population.*

	$\alpha+\frac{K}{N}$	$\alpha+\frac{2K}{N}$	$\alpha+\frac{3K}{N}$	$\alpha+\frac{4K}{N}$	$\alpha+\frac{5K}{N}$	$\alpha+\frac{6K}{N}$
$N=2$	·500	·975	—	—	—	—
$N=3$	·342	·659	·975	—	—	—
$N=4$	·262	·500	·737	·975	—	—
$N=5$	·215	·405	·595	·785	·975	—
$N=10$	·120	·215	·310	·405	·500	·595
$N=25$	·063	·101	·139	·177	·215	·253
$N=50$	·044	·063	·082	·101	·120	·139

(b) *Observations* (as far as ascertainable).

2000 samples of 2	·5	·975	—	—	—	—
1333 " " 3	·34167	·65833	—	—	—	—
1000 " " 4	·26250	·50000	—	—	—	—
1000 " " 5	·215	·405	—	—	—	—
1000 " " 10	·11—·13	·21—·23	·31—·33	—	—	—
1000 " " 25	·06—·07	·10—·11	·13—·14	·17—·18	—	—

In samples of 50 the positions of the expected maxima are becoming close together and are more nearly obscured by the grouping, but they can still be traced in the tables of frequencies or in the histogram of Fig. 10.

(4) One other point remains. Clearly the method of representing theoretically by a Pearson Curve the actual distributions of means fails when N is small but the figures show that the method would appear to commence to be effective from $N = 50$. To obtain verification or otherwise of this the Pearson Curve corresponding to the theoretical values of the betas in the case of $N = 50$ was calculated and fitted to the two distributions obtained by sampling, the ordinary test for Goodness of Fit being applied.

Now N being 50 the theoretical values of ${}_1B_1$ and ${}_1B_2$ are

$${}_1B_1 = 0{\cdot}280{,}332 \text{ and } {}_1B_2 = 3{\cdot}250{,}847$$

whilst those of σ, ν_2 and ν_3 are

$$\sigma = 0{\cdot}29{,}745, \quad \nu_2 = 0{\cdot}088{,}479 \text{ and } \nu_3 = 0{\cdot}014{,}013,$$

the unit being that of the original distribution.

* *Biometrika*, Vol. XIX. p. 244.

The corresponding Pearson Curve is therefore of Type I and its constants are, in the usual notation,

$$r = 34{\cdot}845{,}330, \qquad \epsilon = 182{\cdot}468{,}882,$$
$$b = 0{\cdot}459{,}394, \qquad m_1 = 5{\cdot}418{,}838,$$
$$m_2 = 27{\cdot}426{,}491, \qquad a_1 = 0{\cdot}075{,}791,$$
$$a_2 = 0{\cdot}383{,}603.$$

Taking the total frequency to be 800, $y_0 = 11{,}234{\cdot}0$.

The Mode (which is the working origin) $= 0{\cdot}071{,}497$.

Thus the frequency curve representing the Distribution of Means of Samples of 50 is

$$y = 11{,}234\left(1 + \frac{x}{0{\cdot}075{,}791}\right)^{5{\cdot}418{,}838}\left(1 - \frac{x}{0{\cdot}383{,}603}\right)^{27{\cdot}426{,}491}.$$

This curve is shown in Fig. 10 on p. 385, and its plotted ordinates are:

x (Value of Mean) ...	*0·00725*	*0·01725*	*0·02725*	*0·03725*	*0·04725*	*0·05725*	*0·06725*	*0·07150*
y	29·2	462·9	1939·7	4509·5	7468·2	9982·0	11117·2	11234·0

x (Value of Mean) ...	*0·07725*	*0·08725*	*0·09725*	*0·11725*	*0·13725*	*0·15725*	*0·17725*	*0·19725*
y	11034·0	10020·8	8150·3	4459·4	1909·6	657·5	183·9	41·8

Tests for Goodness of Fit.

The data for these tests were obtained by drawing the curve accurately to a large scale and planimetering the areas corresponding to the frequency groups of the experimental distributions. Of course to fit the experimental distribution of 200 means all curve frequencies were reduced to one quarter.

Before the actual tests are considered it is as well to recall that, as the original distribution consisted of only a few frequency groups, the values of the mean of any sample are discontinuous, being grouped together to prevent the labour of computation becoming impossibly great. We are then attempting to represent this discontinuous result by a continuous frequency curve. This is exemplified by the fact that from the given population as grouped, it is impossible to obtain a sample whose mean is $< 0{\cdot}025$ whilst the curve itself has appreciable frequency nearly down to zero. The upper limit of the value of the mean of a sample is not so definite and the frequencies of curve and distributions corresponding to the higher values of the mean of a sample are quite reasonably similar, so at this end we may safely coalesce the groups in the "tail," as is ordinarily done in Goodness of Fit tests. It is the frequencies corresponding to the lower values of the mean that

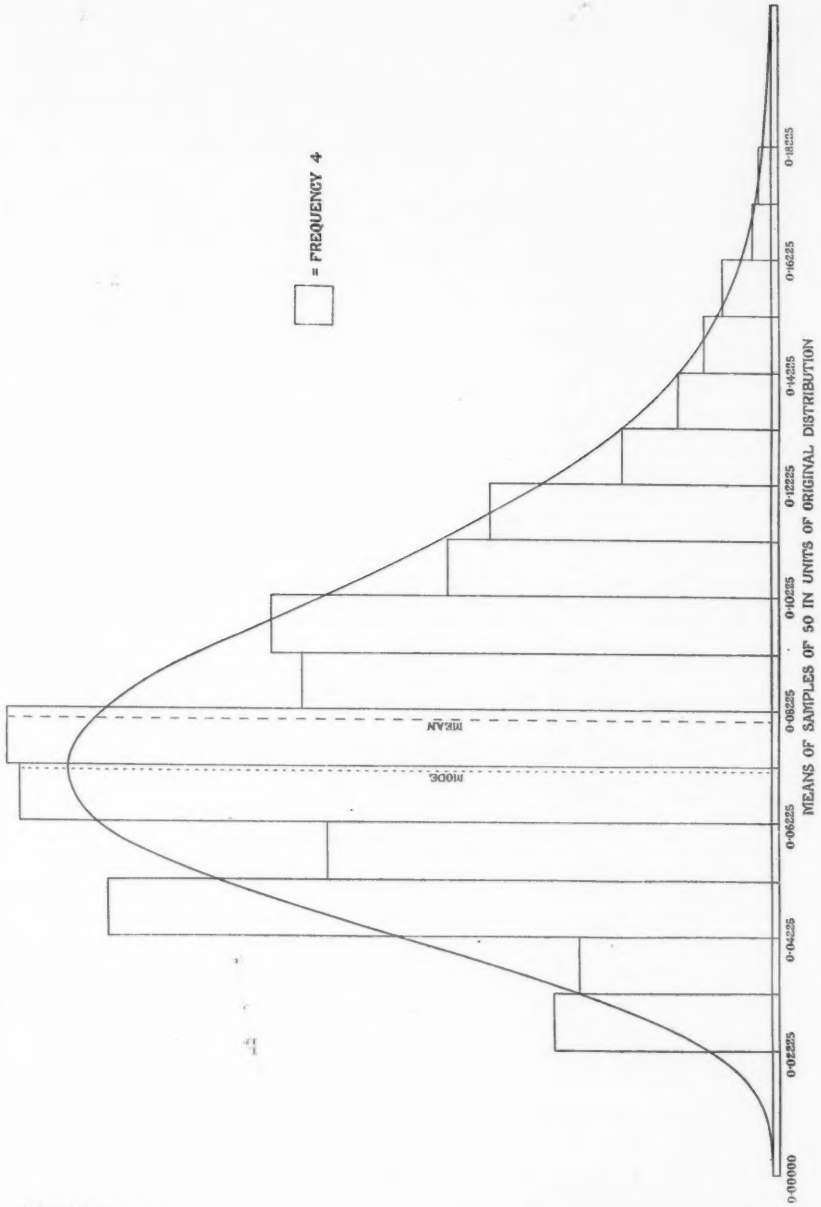

Distribution of Means of 800 Samples of 50 with Fitted Curve from Theoretical Values of β_1 & β_2.
= FREQUENCY 4
MEAN
MODE
0·00000
0·02225
0·04225
0·06225
0·08225
0·10225
0·12225
0·14225
0·16225
0·18225
MEANS OF SAMPLES OF 50 IN UNITS OF ORIGINAL DISTRIBUTION

Fig. 10.

cause the greatest discrepancies between the continuous curve and the discontinuous distributions. Lastly Figs. 9 (*a*) and (*b*) show that even when $N = 50$ the multiple-U form, so apparent in the distributions corresponding to $N < 50$, still persists.

Consider first the fitting of the curve to the distribution of 800 means.

Grouping all frequency beyond 0·15225, as mentioned above, we have 14 groups with $\chi^2 = 41{\cdot}89$, $P = 0{\cdot}0001$.

This is an extremely poor fit, but the large value of χ^2 is easily seen to arise almost entirely from the first four frequency groups, i.e. those corresponding to the low values of the mean. Coalescing these into 2 groups we have with 12 groups $\chi^2 = 11{\cdot}97$, $P = 0{\cdot}3651$, quite a reasonable fit. Then, if we obliterate the last large U form in this distribution by coalescing the two frequency groups corresponding to the ranges 0·08225—0·09225 and 0·09225—0·10225 in the values of the mean we obtain with 11 groups $\chi^2 = 9{\cdot}62$, $P = 0{\cdot}4752$, a still better fit.

Consider now the distribution of 200 means. Grouping all frequency above the value 0·15 of the mean as was done in the previous case, we have with 13 groups $\chi^2 = 21{\cdot}00$, $P = 0{\cdot}0504$, again a distinctly poor fit, whilst the large value of χ^2 is still due to the frequency groups corresponding to the low values of the mean. Coalescing the first two of these frequency groups we have with 12 groups $\chi^2 = 15{\cdot}08$, $P = 0{\cdot}1791$, a better fit if not a very good one. As is to be expected, there being only 200 samples in this distribution, the frequencies corresponding to the large values of the mean are not so even as was the case with the 800 samples and this keeps the value of P down. If we now coalesce the last three groups of this distribution (i.e. the three groups corresponding to the large values of the mean) we obtain with 11 groups $\chi^2 = 11{\cdot}29$, $P = 0{\cdot}3365$, quite a fair fit.

Thus, the original sampled population being U-shaped and of few frequency groups, even when $N = 50$, the discontinuity of the distribution of means is sufficiently marked to make its representation by a continuous curve of very doubtful value; but if such a representation is necessary then there will be a better correspondence between curve and observed distribution, when the frequency groups in the distribution are arranged so that each contains a fair number of the discrete values of the mean.

To obtain further confirmation of this, the distribution of 800 means was subdivided so that the range of values of the mean corresponding to each frequency group was one-half of its previous value.

The distribution obtained was:

Value of Mean ...	*0·02225—0·02725*	*0·02725—*	*0·03225—*	*0·03725—*	*0·04225—*
Frequency ...	18	17	16	15	68

Value of Mean ...	0·04725—	0·05225—	0·05725—	0·06225—	0·06725—	0·07225—
Frequency ...	38	29	42	80	40	41

Value of Mean ...	0·07725—	0·08225—	0·08725—	0·09225—	0·09725—	0·10225—
Frequency ...	81	44	31	36	44	29

Value of Mean ...	0·10725—	0·11225—	0·11725—	0·12225—	0·12725—	0·13225—
Frequency ...	23	23	22	19	5	7

Value of Mean ...	0·13725—	0·14225—	0·14725—	0·15225—	0·15725—	0·16225—
Frequency ...	8	3	8	4	4	1

Value of Mean ...	0·16725—	0·17225—	0·17725—0·18225
Frequency ...	2	1	1

of which, owing to the difficulty of reproduction, no figure is given.

Grouping all frequencies above 0·15225 as before, we have with 27 groups $\chi^2 = 109{\cdot}27$, $P = 0{\cdot}0000$, no fit at all. An inspection of the distribution above, shows that the increase in the number of frequency groups merely accentuates the multiple-U form of the distribution making its representation by a Type I curve quite impossible.

Summary of Results.

(1) When distributions of means in samples of N are obtained by actual sampling from a U-shaped population their representation by a simple continuous curve seems to be quite unsatisfactory until the size of the sample N is at least of the order of 50. This is not due to any extent to the irregularities arising from the grouping of the sampled population, but mainly to the fact that when N is quite small, these distributions are essentially composite in form, because the sampled population is of limited range.

(2) In obtaining sets of samples of N by use of Tippett's *Random Sampling Numbers* it is advisable, in order that the sets may be as random as possible, to

make a fresh allocation of the numbers for each set of samples obtained. The regrouping of the samples in any one set to form a set corresponding to a larger value of N tends to produce a bias in the set so formed.

(3) When such a distribution of means of samples of N begins to be representable by a Pearson Curve derived from the theoretical betas corresponding to it, the goodness of the fit appears to depend more on the way the actual distribution of discontinuous means is grouped into frequency groups than on the actual number of samples taken.

In conclusion the authors desire to express their thanks to Miss Ida McLearn for preparing the diagrams of this paper for publication.

DISCUSSION OF SMALL SAMPLES DRAWN FROM AN INFINITE SKEW POPULATION.

By "SOPHISTER."

1. *Introductory.*

The objects of this paper are to determine:

(1) How far considerable skewness in the population sampled influences the distribution of the constants of small samples.

(2) As far as possible equations describing the distribution of variance, standard deviation and "Student's" z in the case of such small samples.

(3) How far skewness, often unrecognisable in practical testing, may be a source of error when inferences are drawn on the hypothesis that the sampled population follows a normal curve of frequency.

The actual distributions of mean, standard deviation and of

$$z = \frac{\text{Sample mean} - \text{Population mean}}{\text{Sample standard deviation}}$$

are known in the case of sampling from a normal population.

The actual distribution of the means of samples is known in the case of a skew curve of Type III. The statistical constants β_1 and β_2, as well as the mean and standard deviation, are known in the case of means of samples from any population. But the frequency curve for means of samples in any cases but those cited is unknown, although very fair approximations to the distribution of the means of samples can be obtained by using the β_1 and β_2 to compute a Pearson Type curve, if the sampled population consists of a uni-modal curve. This latter consideration must be borne in mind, as is shown by Holzinger and Church in this volume. It may thus be said that the distribution of the mean in small samples for such cases is fairly completely worked out and our only object in the present investigation will be to test into what error we fall, if we assume the distribution to be due to sampling from a normal curve, i.e. assume β_1 and β_2 to be zero and three respectively. This is not done by way of improving the result by actual use of the true values of β_1 and β_2, for in most practical cases we do not know the β_1 and β_2 of the original population, but to indicate what is the extent of error we are liable to make if the original sampled population were really skew, while we have assumed it to be normal.

Turning to the distribution of the variance in samples we shall see that it is possible to find for it—and therefore for the standard deviation—frequency curves which give exceedingly good results, in accordance with the observed frequencies, when we know the constants of the sampled population. These curves enable us to judge how far we may be misled if, when knowing nothing about the original population, we assume it to be normal.

Lastly we come to the distribution of z in small samples. The form of this is known only when the sampled population is normal. We show in this paper that if we could determine the β_1 of the distribution of z in samples, in terms of the constants of the original sampled population, we could obtain a reasonable equation for the frequency curve of z. But so far attempts to obtain this have not been successful. On the other hand the experimental distributions of z in samples from skew populations appear to differ considerably from the theoretical distribution of z on the supposition that the original population sampled has a normal distribution. One has therefore to be careful in drawing inferences from z if there be any suspicion that there is skewness in the sampled population.

The principal previous experimental work on this topic is that of Church*, who does not however investigate the distribution of z for his samples, although such distribution has been recently determined for some of Church's data by Neyman and Pearson†.

In the present paper, in order to emphasise the effect of want of normality in the sampled population, we have used considerably higher values of β_1 and β_2 than those adopted by Church.

Population Sampled.

The population was obtained from a Pearson Type III curve

$$y = y_0 e^{-x} x^7.$$

Samples were drawn using Tippett's Table of Random Sampling Numbers‡ in the manner described by Church§. The 10,000 random numbers of the population were arranged in 41 groups by means of the *Tables of the Incomplete* Γ*-function*, taking ·2 of the standard deviation as a group unit. As we cannot sample from the true curve but only from grouped data, we find on taking the moments that these are altered to some extent.

Constants of Population.

	Constants of Curve	Constants found from Grouped Data	
		Using Sheppard's Corrections	Without Sheppard's Corrections
Mean ...	2·828,43	2·828,44	2·828,44
σ ...	1·0000	·999,857	1·001,522
β_1 ...	·5000	·497,025	·492,087
β_2 ...	3·7500	3·729,789	3·724,935
Skewness ...	+ ·3536	+ ·3536	+ ·3536

* *Biometrika*, Vol. XVIII. p. 321.
† *Biometrika*, Vol. XX A. p. 198.
‡ *Tracts for Computers*, No. XV.
§ *Biometrika*, Vol. XVIII. p. 332.

1000 samples of 5 and 1000 samples of 20 were drawn. For the unit of scale the σ of the sampled population will be used throughout. As the higher β's of the population will be used later, without Sheppard's corrections, they were calculated from the usual relationship with β_1 and β_2* to give

$$\beta_3 = 5{\cdot}608{,}458, \quad \beta_4 = 32{\cdot}418{,}438, \quad \beta_6 = 504{\cdot}365{,}310.$$

2. *Distribution of Means of Samples.*

These are given in Table I and in Diagrams I and II (see pp. 392—3). Now if

Σ, B_1, B_2 are the constants of an infinite population,

σ, β_1, β_2 „ „ a sample of size n,

we know that for random samples

$$\sigma = \frac{\Sigma}{\sqrt{n}}, \quad \beta_1 = \frac{B_1}{n}, \quad \beta_2 = \frac{1}{n}(B_2 - 3) + 3.$$

Comparing these values with those obtained from the distributions, we have:

Constants of Distributions of Means.

	$n=20$		$n=5$	
	Actual Values	Theoretical	Actual Values	Theoretical
Mean ...	2·8236	2·8284 ± ·0048	2·8102	2·8284 ± ·0095
σ ...	·2291	·2236 ± ·0034	·4304	·4471 ± ·0070
β_1 ...	·0216	·0249 [± ·0168]	·0281	·0994 ± ·0339
β_2 ...	3·0800	3·0365 ± ·1175	2·8391	3·1460 ± ·1544

Both these distributions are fitted with normal curves in Table I, using the mean of the population and $\sigma = \Sigma/\sqrt{n}$, giving $P = {\cdot}759$ for 25 groups in the first case and $P = {\cdot}182$ for 25 groups in the second case. This shows that even for a population of considerable skewness the normal theory represents the distribution of the means of our samples satisfactorily.

The distribution of means of samples of 20 is so nearly normal, with $\beta_1 = {\cdot}0216$ and $\beta_2 = 3{\cdot}0800$, that nothing would be gained by fitting the correct Type III curve. With samples of 5, the Type III curve has been fitted in Table I. Since $\beta_1 = {\cdot}0994$, and for this curve $p = \dfrac{4}{\beta_1} - 1 = 39{\cdot}239$, we can fit the curve by interpolation from the *Tables of the Incomplete* Γ*-function* for this value of p. The goodness of fit test gives $P = {\cdot}284$ for 24 groups which is only slightly better than the fit of the normal curve.

We must reach the conclusion, accordingly, that we should not be able to form from the distributions of the means of our samples any accurate opinion as to whether the population sampled, if unknown, differed from a normal population.

* *Tables for Statisticians*, p. lxi.

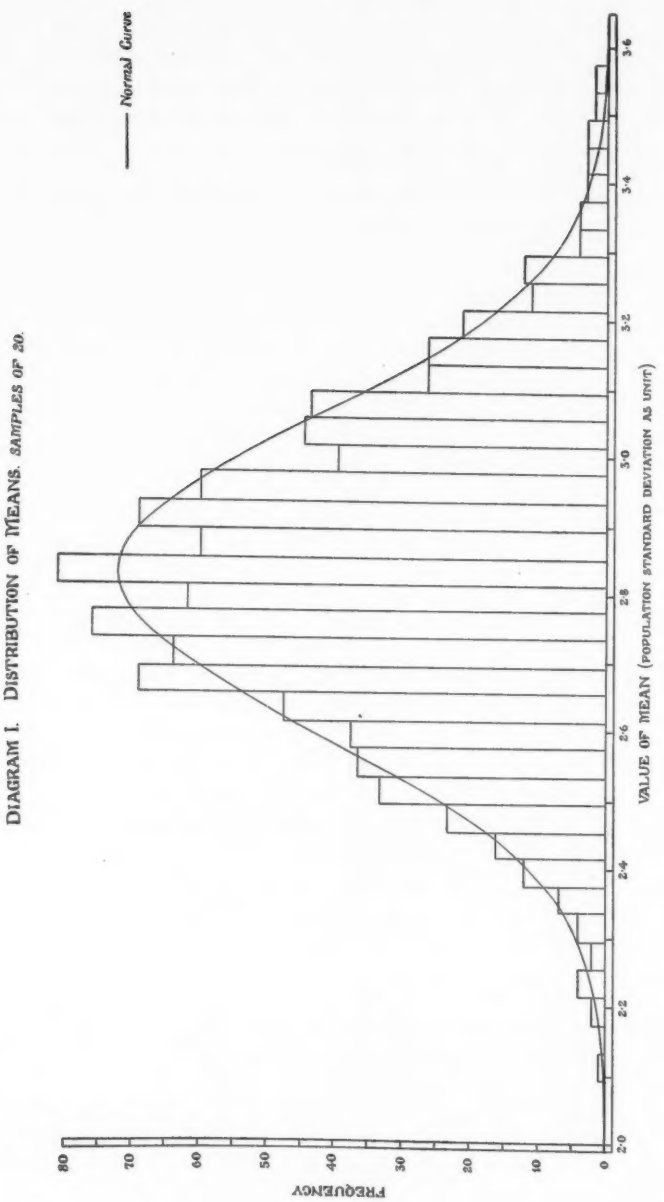


DIAGRAM I. DISTRIBUTION OF MEANS. *SAMPLES OF 20.*

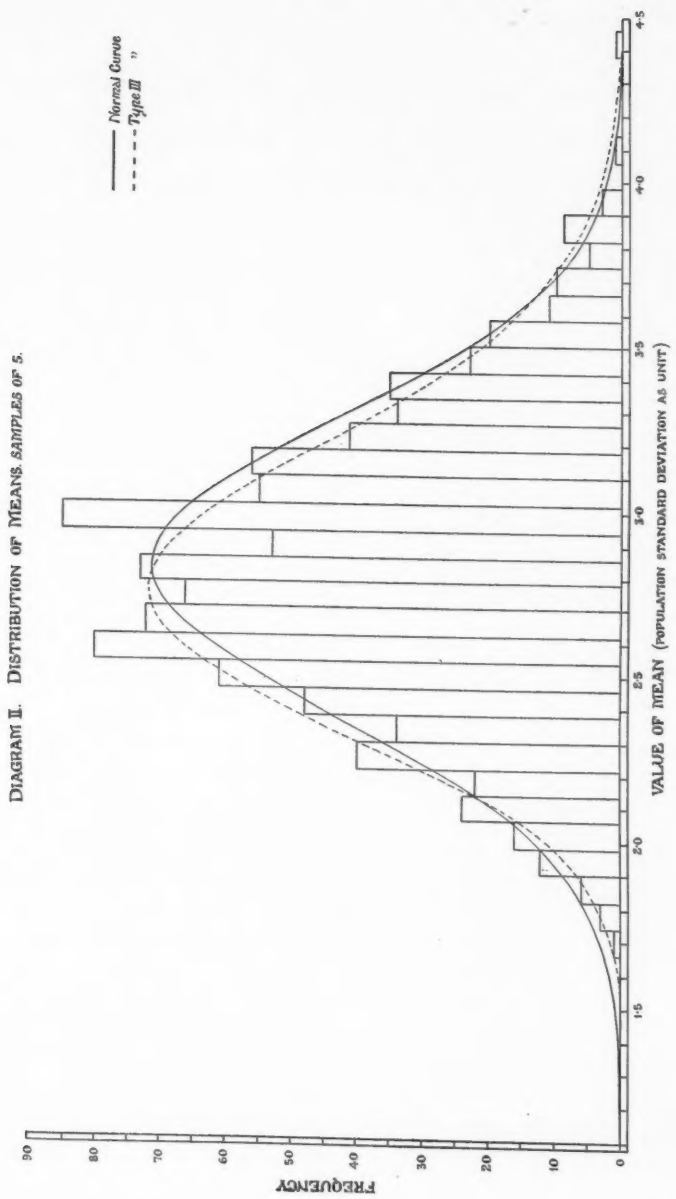


DIAGRAM II. DISTRIBUTION OF MEANS. *SAMPLES OF 5.*

TABLE I.

Means of Samples.

$n=20$			$n=5$			
Range of Values of Mean	Observed	Normal Theory	Range of Values of Mean	Observed	Normal Theory	Type III Curve
2·095—2·135	1		*1·66—1·74*	1		
2·135—2·175	0		*1·74—1·82*	3	18·93	11·25
2·175—2·215	2	13·66	*1·82—1·90*	6		
2·215—2·255	4		*1·90—1·98*	12	9·96	22·60
2·255—2·295	2		*1·98—2·06*	16	13·96	
2·295—2·335	4		*2·06—2·14*	24	18·98	19·77
2·335—2·375	7	18·55	*2·14—2·22*	22	24·97	27·18
2·375—2·415	12		*2·22—2·30*	40	31·84	35·96
2·415—2·455	16	15·22	*2·30—2·38*	34	39·32	44·66
2·455—2·495	23	20·50	*2·38—2·46*	48	47·01	53·04
2·495—2·535	33	26·75	*2·46—2·54*	61	54·46	60·56
2·535—2·575	36	33·80	*2·54—2·62*	80	61·11	66·50
2·575—2·615	37	41·39	*2·62—2·70*	72	66·44	70·58
2·615—2·655	47	49·07	*2·70—2·78*	66	69·90	71·86
2·655—2·695	68	56·38	*2·78—2·86*	73	71·26	71·14
2·695—2·735	63	62·69	*2·86—2·94*	53	70·37	68·17
2·735—2·775	75	67·54	*2·94—3·02*	85	67·31	63·43
2·775—2·815	61	70·49	*3·02—3·10*	55	62·35	57·19
2·815—2·855	80	71·24	*3·10—3·18*	56	55·95	50·52
2·855—2·895	59	69·75	*3·18—3·26*	41	48·63	43·45
2·895—2·935	68	66·14	*3·26—3·34*	34	40·94	36·60
2·935—2·975	59	60·75	*3·34—3·42*	35	33·39	29·73
2·975—3·015	39	54·05	*3·42—3·50*	23	26·36	23·77
3·015—3·055	44	46·59	*3·50—3·58*	20	20·16	18·93
3·055—3·095	43	38·87	*3·58—3·66*	11	14·94	14·56
3·095—3·135	26	31·41	*3·66—3·74*	10	10·71	11·01
3·135—3·175	26	24·60	*3·74—3·82*	5		
3·175—3·215	21	18·65	*3·82—3·90*	9		
3·215—3·255	11	13·71	*3·90—3·98*	3		
3·255—3·295	12	16·46	*3·98—4·06*	—		
3·295—3·335	4		*4·06—4·14*	1	20·75	27·54
3·335—3·375	4		*4·14—4·22*	—		
3·375—3·415	3		*4·22—4·30*	—		
3·415—3·455	3	11·74	*4·30—4·38*	—		
3·455—3·495	3		*4·38—4·46*	1		
3·495—3·535	2					
3·535—3·575	2					
χ^2		18·868	χ^2		30·709	26·372
P		·759	P		·182	·284

3. *Distribution of Ranges of Samples.*

The distribution of ranges of samples from a normal population has been studied by Tippett*, and Dr E. S. Pearson† has collected the results in a convenient

* *Biometrika*, Vol. XVII. p. 364.

† *Biometrika*, Vol. XVIII. p. 192.

table. The constants of the distributions in Table II are compared with the normal theory below:

Constants of Distributions of Range.

	$n=20$		$n=5$	
	Actual Values	Normal Theory	Actual Values	Normal Theory
Mean ...	3·7594	3·7349 ± ·0155	2·2920	2·3259 ± ·0184
σ ...	·8309	·729 ± ·0117	·8935	·8641 ± ·0136
β_1 ...	·5353	·161 ± ·047	·2299	·2167 ± ·0499
β_2 ...	3·9122	3·26 ± ·19	3·0523	3·1693 ± ·1621

The equations for the curves for the normal theory are given by "Student*." These have been fitted to the distributions in Table II and the best fitting curves are shown for comparison: in each case the mid-ordinates of the groups were computed and corrections applied to give frequencies. The constants of the curves are:

$n=20$			$n=5$		
Type VI. $y=y_0(x-a)^{q_2}x^{-q_1}$			Type I. $y=y_0\left(1+\frac{x}{a_1}\right)^{m_1}\left(1-\frac{x}{a_2}\right)^{m_2}$		
	Best Fitting Curve	Normal Theory		Best Fitting Curve	Normal Theory
$\log y_0$ ...	93·822,599	603·835,959	y_0 ...	443·5695	465·255
a ...	13·8632	41·567	a_1 ...	2·0604	2·478
q_1 ...	78·8732	373·818	a_2 ...	7·0226	10·892
q_2 ...	11·6207	31·440	m_1 ...	3·7851	6·601
			m_2 ...	12·9010	29·009
Mean ...	3·7594	3·7349	Mean ...	2·2920	2·3259
Mode ...	3·4735	3·5904	Mode ...	2·0264	2·1022
Origin ...	−12·7851	−41·7936	Origin ...	At mode	At mode
Start ...	1·0781	− ·2266	Start ...	− ·0340	− ·3758
Upper limit ...	∞	∞	Upper limit ...	9·0490	12·9942

The curve for the normal theory fits the smaller sample well, giving $P=·824$ as against $P=·994$ for the best fitting curve, both for 20 groups. With the larger sample, however, the normal theory only gives $P=·011$ for 18 groups as against $P=·914$ for 20 groups of the best fitting curve.

The curves are shown in Diagrams III and IV (see pp. 397—8).

4. *Distributions of Variance in Samples.*

These are given in Table III (see p. 407). The constants of the distribution of samples of 5 are corrected for abruptness by subdividing the first group† as shown.

* *Biometrika*, Vol. XIX. p. 163.

† *Biometrika*, Vol. XII. p. 231.

TABLE II.

Distribution of Range of Samples.

$n=20$				$n=5$			
Central Value of Range Group	Observed	Best Fitting Curve	Normal Theory	Central Value of Range Group	Observed	Best Fitting Curve	Normal Theory
				(Less than ·3)	—	15·7	12·3
				·4	3		
				·6	13		
				·8	26	23·2	18·1
				1·0	36	38·0	31·4
				1·2	53	53·5	47·1
				1·4	68	67·6	63·1
				1·6	75	78·7	77·0
(Less than 2·0)	—	13·3	13·3	1·8	89	85·8	87·0
2·0	7			2·0	81	88·5	92·2
2·2	11			2·2	94	87·1	92·3
2·4	19	22·6	18·2	2·4	81	82·2	87·9
2·6	32	41·2	33·6	2·6	75	74·8	80·1
2·8	64	62·1	53·7	2·8	72	65·8	70·0
3·0	84	81·2	75·4	3·0	53	56·0	59·1
3·2	103	95·1	95·9	3·2	45	46·2	48·0
3·4	100	101·6	106·8	3·4	39	37·0	37·8
3·6	99	100·8	110·6	3·6	29	28·9	28·9
3·8	96	94·0	106·3	3·8	15	21·8	21·4
4·0	78	83·3	95·1	4·0	18	16·0	15·4
4·2	79	70·8	80·0	4·2	11	19·3	18·1
4·4	50	57·8	63·4	4·4	9		
4·6	51	45·9	47·7	4·6	6		
4·8	32	35·4	34·3	4·8	5		
5·0	32	26·7	23·6	5·0	2		
5·2	15	19·7	15·6	5·2	1	13·9	12·8
5·4	15	14·3	9·9	5·4	—		
5·6	9	10·3		5·6	—		
5·8	10	12·3		5·8	1		
6·0	4			(Over 5·9)	—		
6·2	3						
6·4	—		16·6				
6·6	2						
6·8	1	11·6					
7·0	3						
7·2	—						
7·4	1						
(Over 7·5)	—						
Number of Groups		20	18	Number of Groups		20	20
χ^2		11·25	36·04	χ^2		5·325	12·574
P		·914	·011	P		·994	·859

The corrections, however, were small and almost unnecessary. The distribution of samples of 20 is treated with the usual Sheppard's corrections.

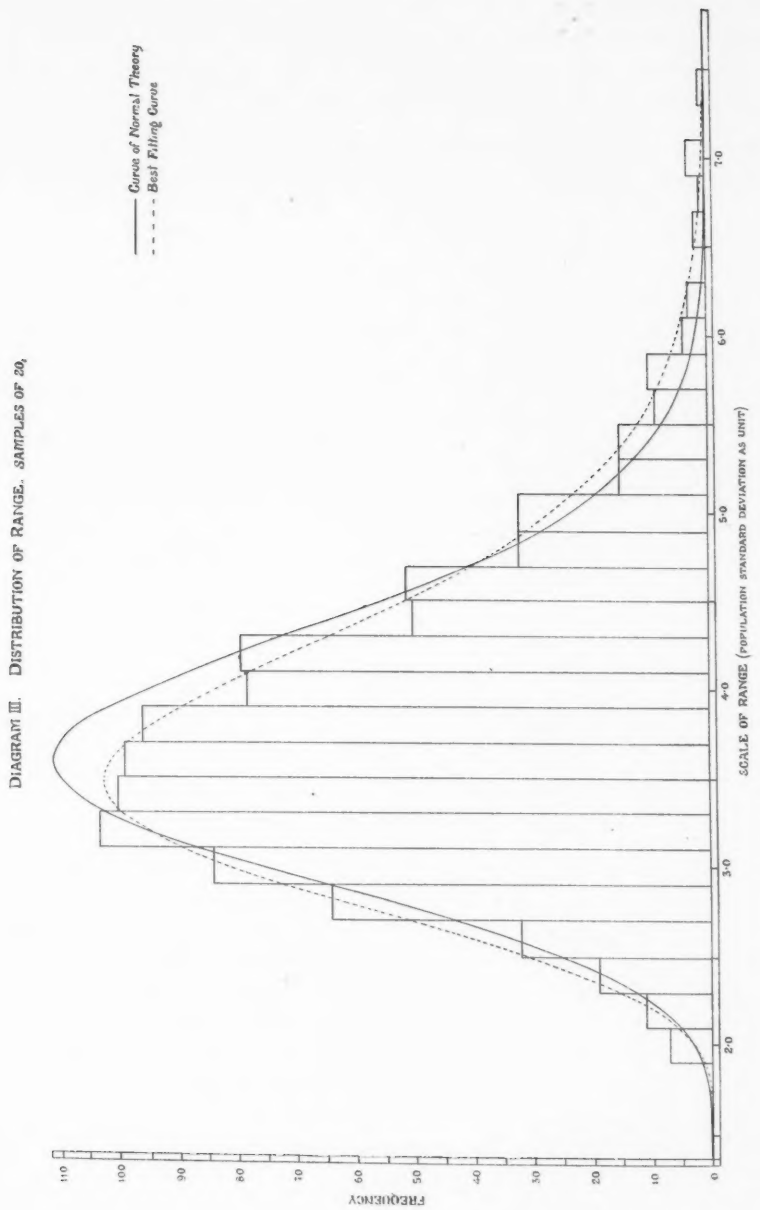


DIAGRAM III. DISTRIBUTION OF RANGE. *SAMPLES OF 20.*

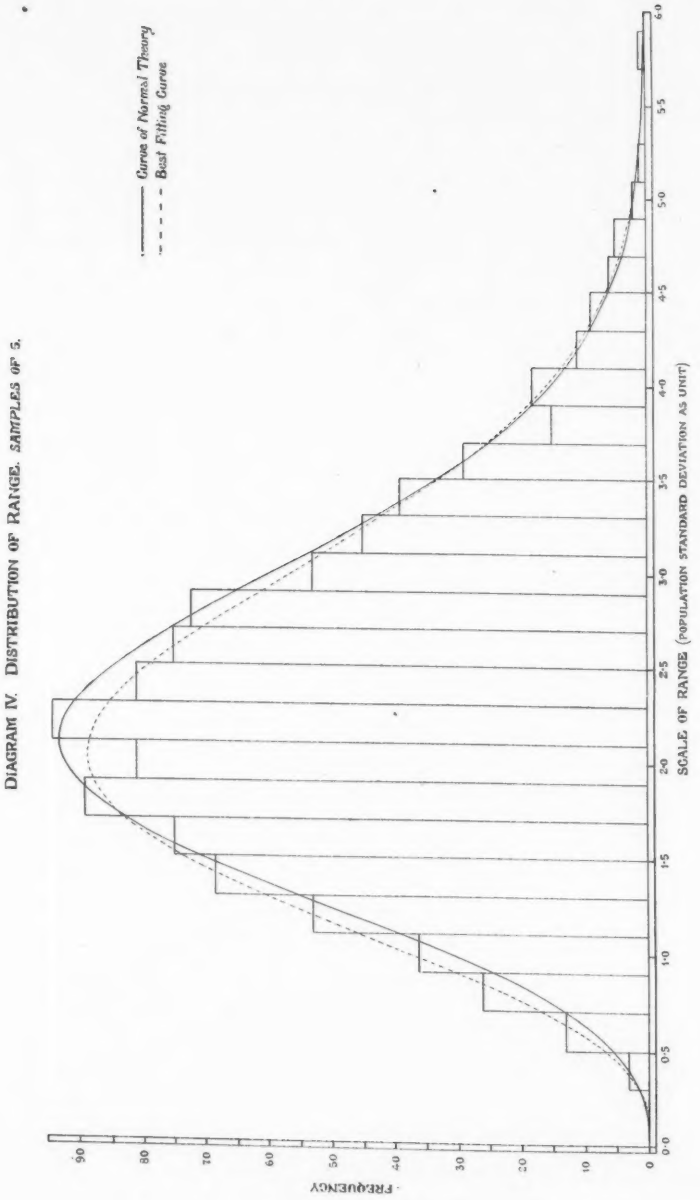


DIAGRAM IV. DISTRIBUTION OF RANGE. *SAMPLES OF 5.*

Now the β relations for variance in samples have been worked out by Tchouproff* and Church †. Following the method of the latter, we obtain the constants of the distributions from the higher β's of the population, up to β_6, the values calculated without the use of Sheppard's corrections being as follows:

Constants of Distributions of Variance.

	$n=20$		$n=5$	
	Actual Values	Church's Equations	Actual Values	Church's Equations
Mean ...	4·8340	4·7645	4·0007	4·0122
σ ...	1·8228	1·7923	3·0417	3·2222
β_1 ...	1·3812	1·0298	2·4654	4·2966
β_2 ...	6·0846	4·9268	6·3859	10·8571
Skewness ...	+0·4064	+0·4254	+0·9055	+ 0·7363
k ...	+0·6967	+1·2737	−4·7933	2·3913
Type ...	IV	VI	I	VI

The Type VI curves $y = y_0 (x-a)^{q_2} x^{-q_1}$ corresponding to the theoretical constants were calculated and fitted to the data by computing the mid-ordinates and applying the proper corrections, with the following result:

	$n=20$	$n=5$
q_1	36·903,611	14·649,447
q_2	12·158,795	·836,869
a	8·745,639	24·995,943
log y_0	35·864,4750	22·996,1786
Origin	− 9·0409	−24·8707
Start of Curve ...	− 0·2953	0·1253
Mode	4·0021	1·6397
No. of Groups ...	20	21
χ^2	27·689	17·43
P	·09	·62

The fit of both of these is satisfactory. The rather low values for samples of 20 can be explained by the irregularities of random sampling. The curves are illustrated in Diagrams V and VI (see pp. 400—1).

Professor Karl Pearson has very kindly suggested a method of obtaining a curve to represent the sampling distribution of variance, which leads to a satisfactory agreement in the two cases of the present sampling and in that of Church.

5. *An approximate Method of finding the Distribution of the Variance of Samples.*

Except in very special cases the problem of the distribution of variance (σ^2) of samples has not so far been solved, and the chance of solving it otherwise than

* *Biometrika*, Vol. XII. p. 193.

† *Biometrika*, Vol. XVII. p. 79 and Vol. XVIII. p. 321.

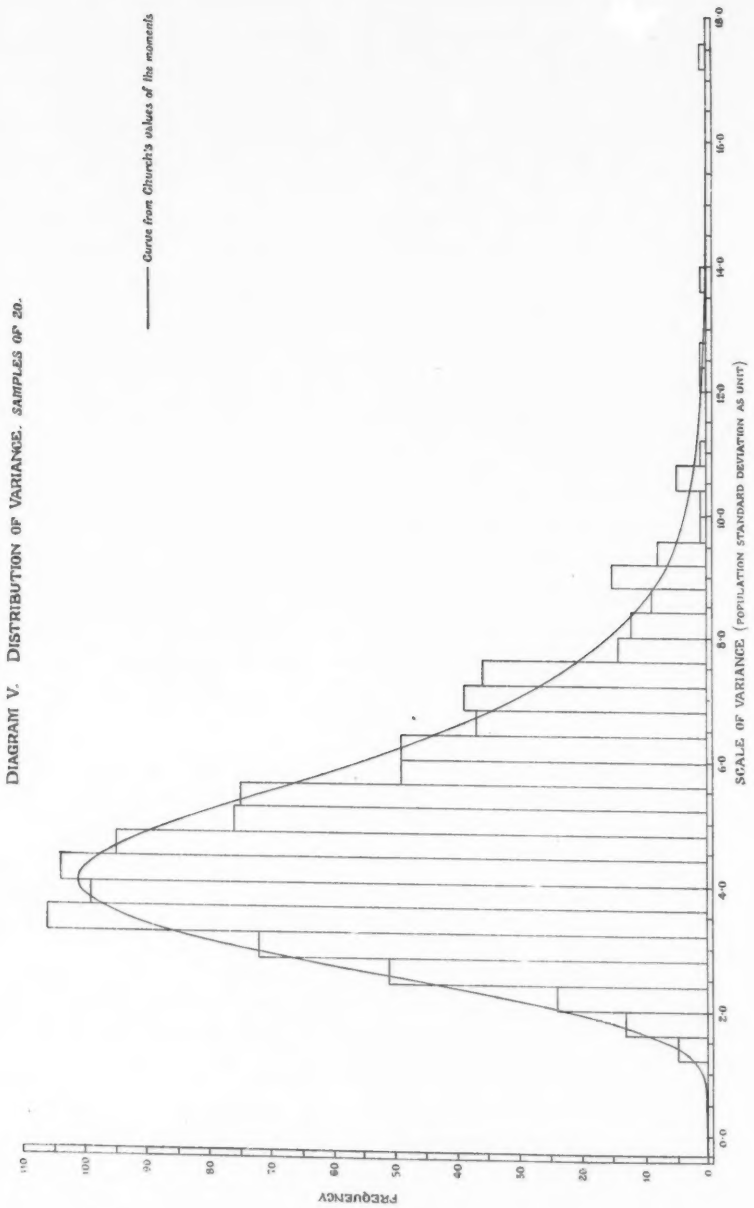


DIAGRAM V. DISTRIBUTION OF VARIANCE. SAMPLES OF 20.
—— Curve from Church's values of the moments
FREQUENCY
0 10 20 30 40 50 60 70 80 90 100 110
SCALE OF VARIANCE (POPULATION STANDARD DEVIATION AS UNIT)
0·0 2·0 4·0 6·0 8·0 10·0 12·0 14·0 16·0 18·0

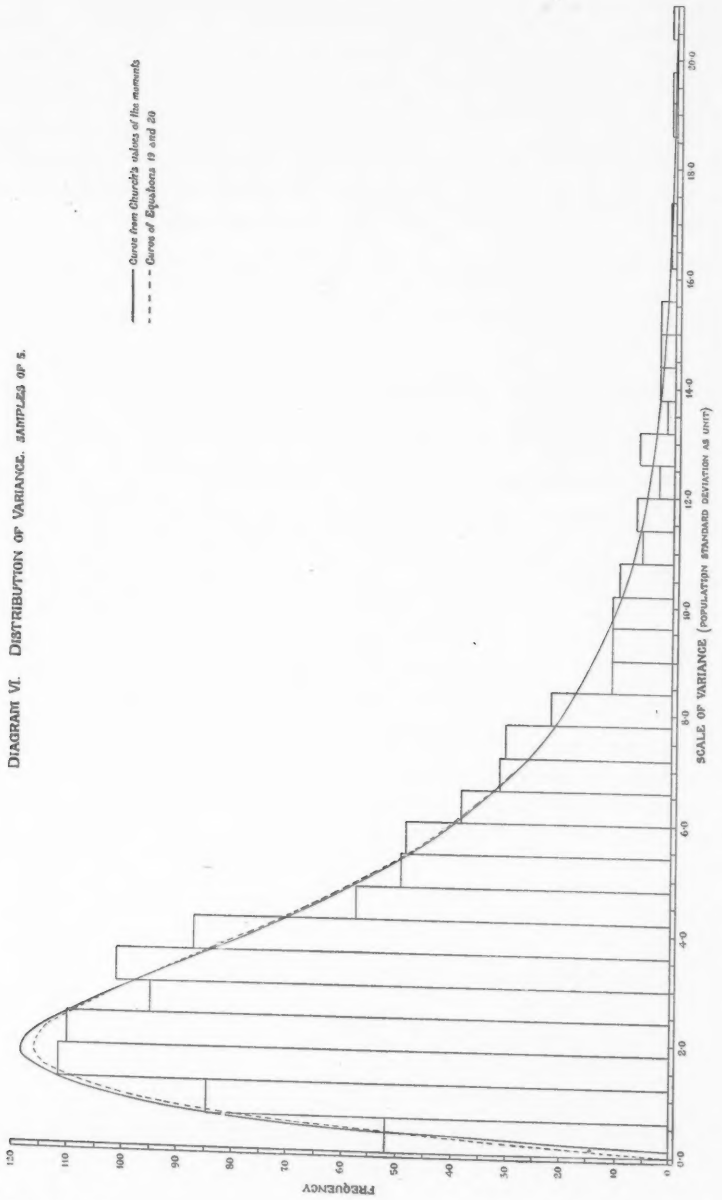


DIAGRAM VI. DISTRIBUTION OF VARIANCE. SAMPLES OF 5.
Curve from Church's values of the moments
Curve of Equations 19 and 20
SCALE OF VARIANCE (POPULATION STANDARD DEVIATION AS UNIT)
FREQUENCY

approximately seems small*. Several sampling experiments, however, appear to indicate that with considerable degrees of skewness in the sampled population, a reasonably close approximation to the frequency distribution of the variance can be obtained in the manner indicated below. It is clear in the first place that if the frequency distribution in the sampled population be not limited at both ends of the range, the frequency curve for σ^2 will run from zero to infinity theoretically. Further, in the case of a frequency curve of finite range, if there be considerable skewness, the variance curve for samples will run from zero to a very considerable distance.

We shall suppose the sampling to take place from an infinitely large population, of which the standard deviation is Σ, and the β-coefficients are B_1 and B_2. We will take the size of the sample to be n and $\mu_2\,(\sigma^2)$ to represent the variance in any sample.

Then the mean value† will be given by

$$\bar{\mu}_2 = \frac{n-1}{n}\,\Sigma^2 \quad \text{.....(1),}$$

and the variance of the variances by

$$\sigma^2_{\mu_2} = \frac{(n-1)^2}{n^3}\,\Sigma^4\left(B_2 - 3 + \frac{2n}{n-1}\right) \quad \text{.....(2).}$$

If we assume the frequency curve of variances to be limited at one end only of the range, then curves of Type III, Type V and Type VI are possible distributions. But as Types III and V are only limiting cases of Type VI it will be sufficient to consider Type VI as a possible curve for the distribution of variances in samples. Thus far in sampling experiments—and with a considerable range in size of samples—Type VI has repeatedly occurred. Even when the β's of the observed samples fell into the area of Type IV it has been found that the β's of the variance curve, as calculated from Church's values of the moments of the variance distribution, lie within the Type VI area, and the observed β's lying in the Type IV area only differed from the theoretical β's by quantities of the order of the probable errors, which in this neighbourhood may be considerable. Curves, however, of Type VI fitted from the theoretical β's gave quite reasonably good fits to the observations.

Accordingly we shall start with the assumption that the distribution of variance in samples is given by a Type VI curve of the form

$$y = y_0\,\mu_2^{q_2}/(\mu_2 + a)^{q_1},$$

with the range from $\mu_2 = 0$ to $\mu_2 = \infty$.

The constant y_0 need not trouble us, as it will be determined by the number of samples taken. It remains therefore to determine the constants q_1, q_2 and a.

* Cf. E. S. Pearson, *Biometrika*, Vol. XIX. p. 245.

† *Biometrika*, Vol. XVII. p. 477.

Now let the first three moment coefficients of μ_2 in samples be

$$\text{Mean} = {}_2M_1',$$
$$\sigma^2_{\mu_2} = {}_2M_2,$$
$$\text{Third moment coefficient} = {}_2M_3.$$

These values are known. Church's value for ${}_2M_3$ can be put into a form simpler for calculation as below. Thus

$${}_2M_1' = \bar{\mu}_2 = \Sigma^2\left(1-\frac{1}{n}\right) \quad \text{.....(1) } bis,$$

$${}_2M_2 = \frac{2\Sigma^4}{n}\left(1-\frac{1}{n}\right)\left\{1+\tfrac{1}{2}(B_2-3)\left(1-\frac{1}{n}\right)\right\} \quad \text{.....(2) } bis,$$

and

$${}_2M_3 = \frac{\Sigma^6}{n^2}\left(1-\frac{1}{n}\right)\left\{B_4\left(1-\frac{1}{n}\right)^2 - 3B_2\left(1-\frac{1}{n}\right)\left(1-\frac{5}{n}\right) - 2B_1\left(3-\frac{6}{n}+\frac{5}{n^2}\right)+2\left(1-\frac{12}{n}+\frac{15}{n^2}\right)\right\} \quad \text{.....(3).}$$

This may be conveniently arranged as

$${}_2M_3 = \frac{\Sigma^6}{n^2}\left(1-\frac{1}{n}\right)\left\{(B_4-15)\left(1-\frac{1}{n}\right)^2 - 3(B_2-3)\left(1-\frac{1}{n}\right)\left(1-\frac{5}{n}\right) - 2B_1\left(3-\frac{6}{n}+\frac{5}{n^2}\right)+8\right\}.$$

Now transfer these to the start of the curve, i.e. to the point at which $\mu_2 = 0$, and write their values ${}_2M_1'$, ${}_2M_2'$, ${}_2M_3'$, so that

$${}_2M_2' = {}_2M_2 + {}_2M_1'^2,$$
$${}_2M_3' = {}_2M_3 + 3\,{}_2M_1'\,{}_2M_2 + {}_2M_1'^3,$$

whence we find

$${}_2M_1' = \Sigma^2\left(1-\frac{1}{n}\right) \quad \text{.....(4),}$$

$${}_2M_2' = \frac{\Sigma^4}{n}\left(1-\frac{1}{n}\right)\left\{n+1+(B_2-3)\left(1-\frac{1}{n}\right)\right\} \quad \text{.....(5),}$$

$${}_2M_3' = \frac{\Sigma^6}{n^2}\left(1-\frac{1}{n}\right)\left\{n^2\left(1-\frac{1}{n}\right)\left(1+\frac{5}{n}\right)+8+(B_4-15)\left(1-\frac{1}{n}\right)^2 + 3(B_2-3)\left(1-\frac{1}{n}\right)\left(n-2+\frac{5}{n}\right) - 2B_1\left(3-\frac{6}{n}+\frac{5}{n^2}\right)\right\} \quad \text{.....(6),}$$

or

$${}_2M_3' = \frac{\Sigma^6}{n^2}\left(1-\frac{1}{n}\right)\left\{(n+1)(n+3)+(B_4-15)\left(1-\frac{1}{n}\right)^2 + 3(B_2-3)\left(1-\frac{1}{n}\right)\left(n-2+\frac{5}{n}\right) - 2B_1\left(3-\frac{6}{n}+\frac{5}{n^2}\right)\right\} \quad \text{.....(6) } bis.$$

We will now write

$$\gamma_1 = \frac{{}_2M_2'}{{}_2M_1'^2} \quad \text{and} \quad \gamma_2 = \frac{{}_2M_3'}{{}_2M_2'\,{}_2M_1'},$$

and from (4), (5) and (6) *bis* deduce

$$\gamma_1 = \frac{1+\frac{1}{n}}{1-\frac{1}{n}} + \frac{1}{n}(B_2-3) \quad \text{.....(7),}$$

$$\gamma_2 = \frac{\left(1+\frac{1}{n}\right)\left(1+\frac{3}{n}\right)+\frac{1}{n^2}\left(1-\frac{1}{n}\right)^2(B_1-15)+\frac{3}{n}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}+\frac{5}{n^2}\right)(B_2-3)-\frac{2}{n^2}\left(3-\frac{6}{n}+\frac{5}{n^2}\right)B_1}{\left(1-\frac{1}{n}\right)^2\left\{\frac{1+\frac{1}{n}}{1-\frac{1}{n}}+\frac{1}{n}(B_2-3)\right\}}$$

.....(8).

Now returning to the equation to our curve $y = y_0 \frac{{\mu_2}^{q_2}}{(\mu_2+a)^{q_1}}$,

$$N \times {}_2M_s' = y_0 \int_0^\infty \mu_2^{q_2+s}/(\mu_2+a)^{q_1}\, d\mu_2; \text{ and taking } \mu_2 = a\left(\frac{1}{u}-1\right)$$

$$= y_0 \int_0^1 a^{s+q_2+1-q_1}(1-u)^{s+q_2}\, u^{q_1-s-q_2-2}\, du$$

$$= y_0\, a^{s+q_2+1-q_1}\, B(s+q_2+1,\ q_1-q_2-s-1).$$

Accordingly
$$y_0 = \frac{N}{a^{q_2+1-q_1}B(q_2+1,\ q_1-q_2-1)} \quad \text{.....(9),}$$

which determines y_0, when N the number of samples is known. Generally

$${}_2M_s' = a^s \frac{B(s+q_2+1,\ q_1-q_2-s-1)}{B(q_2+1,\ q_1-q_2-1)}.$$

Putting $s=1$,
$${}_2M_1' = \Sigma^2\left(1-\frac{1}{n}\right) = \frac{a(q_2+1)}{q_1-q_2-2}$$

or
$$a = \Sigma^2\left(1-\frac{1}{n}\right)\frac{q_1-q_2-2}{q_2+1} \quad \text{.....(10).}$$

Similarly
$${}_2M_2' = \frac{a^2(q_2+2)(q_2+1)}{(q_1-q_2-2)(q_1-q_2-3)} \quad \text{.....(11),}$$

and
$${}_2M_3' = \frac{a^3(q_2+3)(q_2+2)(q_2+1)}{(q_1-q_2-2)(q_1-q_2-3)(q_1-q_2-4)} \quad \text{.....(12).}$$

Hence
$$\gamma_1 = \frac{(q_2+2)(q_1-q_2-2)}{(q_2+1)(q_1-q_2-3)} \quad \text{.....(13),}$$

$$\gamma_2 = \frac{(q_2+3)(q_1-q_2-2)}{(q_2+1)(q_1-q_2-4)} \quad \text{.....(14).}$$

Whence, solving, we find

$$q_2 = \frac{3\gamma_1+\gamma_1\gamma_2-4\gamma_2}{2\gamma_2-\gamma_1\gamma_2-\gamma_1} \text{ or } \frac{2(\gamma_1-\gamma_2)}{2\gamma_2-\gamma_1\gamma_2-\gamma_1} - 1 \quad \text{.....(15),}$$

$$q_1 = 1+2(\gamma_1-\gamma_2)\left\{\frac{1}{2\gamma_1-\gamma_2-1}+\frac{1}{2\gamma_2-\gamma_1-\gamma_1\gamma_2}\right\} \quad \text{.....(16).}$$

We will now adapt our formulae to the special case of sampling from a curve of Type III, $y = y_0 x^p e^{-\gamma x}$.

The following are the values of the moments about the mean:

$$\mu_2 = \frac{p+1}{\gamma^2}, \quad \mu_3 = \frac{2(p+1)}{\gamma^3}, \quad \mu_4 = \frac{6(p+1)}{\gamma^4} + \frac{3(p+1)^2}{\gamma^4},$$

or

$$\mu_3 = \frac{2}{\sqrt{p+1}}\sigma^3, \quad \mu_4 = \sigma^4\left(3 + \frac{6}{p+1}\right),$$

leading to

$$\beta_1 = \frac{4}{p+1}, \quad \beta_2 - 3 = \frac{6}{p+1} \quad \text{.....(17).}$$

Further

$$\mu_5 = \frac{4\sigma^5}{\sqrt{p+1}}\left(5 + \frac{6}{p+1}\right), \quad \mu_6 = 5\sigma^6\left\{3 + \frac{26}{p+1} + \frac{24}{(p+1)^2}\right\}.$$

Hence

$$\beta_3 = \frac{8}{p+1}\left(5 + \frac{6}{p+1}\right), \quad \beta_4 - 15 = \frac{10}{p+1}\left(13 + \frac{12}{p+1}\right) \text{.....(18).}$$

Applying these to (7) and (8) we have

$$\gamma_1 = \frac{n+1}{n-1} + \frac{6}{n(p+1)} \text{.....(19),}$$

$$\gamma_2 = \frac{(n+1)(n+3) + \left(18n + 52 - \frac{86}{n}\right)\frac{1}{p+1} + 120\left(\frac{n-1}{n}\right)^2\frac{1}{(p+1)^2}}{(n-1)^2\left\{\frac{n+1}{n-1} + \frac{6}{n(p+1)}\right\}} \text{.....(20).}$$

We have now to substitute our special values of n, B_1, B_2 and B_4 in the accurate equations (7) and (8). Also assuming the population to be described by a Type III curve such that $p = \frac{4}{B_1} - 1 = 7{\cdot}128{,}6466$, we can obtain values of γ_1 and γ_2 from equations (19) and (20). We find quite good agreement of the latter with the more accurate figures from (7) and (8).

	$n=20$		$n=5$	
	General Equations (7) and (8)	Type III (19) and (20)	General Equations (7) and (8)	Type III (19) and (20)
γ_1	1·141,5099	1·142,1697	1·644,9869	1·647,6261
γ_2	1·295,2574	1·297,0284	2·436,9049	2·447,2871

The disagreement, such as it is between these values, arises simply from the fact that the grouped distribution actually sampled from is not quite the same as a continuous Type III curve.

Using now the more accurate values of γ_1 and γ_2 to find q_1 and q_2 by equations

(15) and (16), and comparing with the q_1 and q_2 obtained first for the curve fitted, using Church's equations, we have the following results:

	$n=20$		$n=5$	
	Equations (15) and (16)	Church's Equations	Equations (15) and (16)	Church's Equations
q_1	36·535,032	36·903,611	13·810,387	14·649,447
q_2	9·407,965	12·158,795	1·030,940	·836,869

Finally, using the relationships

$$a = \Sigma^2\left(1-\frac{1}{n}\right)\frac{q_1-q_2-2}{q_2+1} \qquad \text{.....(10)}$$

and

$$y_0 = \frac{Na^{q_1-q_2-1}\,\Gamma(q_1)}{\Gamma(q_1-q_2-1)\,\Gamma(q_2+1)},$$

the distributions were fitted with Type VI curves, having our new q_1 and q_2, and the results are shown below:

$$y = y_0\mu_2^{q_2}/(\mu_2+a)^{q_1}.$$

	$n=20$	$n=5$
q_1	36·535,0324	13·810,3869
q_2	9·407,9650	1·030,940
a	11·502,4661	21·2952
log y_0	40·228,7697	20·853,0061
Origin	Zero	Zero
Start of Curve ...	Zero.	Zero
Mode	3·989,12	1·7179
No. of Groups ...	20	21
χ^2	27·257	12·857
P	·10	·88

We have thus obtained curves which fit our data rather better than those derived from the β's of Church's formulae, and have also the advantage of starting from the true point of zero σ^2. By assuming that the curve will be of Type VI, we have also only had to use the first four β's of the sampled population. The curve for $n=5$ is shown in Diagram VI (see p. 401), that for $n=20$ coincides so closely with the previous curve from Church's equations (Diagram V, p. 400) that it could not be shown separately.

6. *Distribution of Standard Deviations of Samples.*

An element of the distribution curve of μ_2 is $y\,d\mu_2$ and the size of this must be the same for σ and for μ_2, or

$$y\,d\mu_2 = y_0\mu_2^{q_2}/(\mu_2+a)^{q_1}\,d\mu_2$$

$$= 2y_0\frac{\sigma^{2q_2+1}}{(\sigma^2+a)^{q_1}}\,d\sigma,$$

TABLE III.

Variance of Samples.

$n=20$				$n=5$			
Value of σ^2	Observed	Church's Equations	Equations (19) and (20)	Value of σ^2	Observed	Church's Equations	Equations (19) and (20)
1·2— 1·6	5			*0—·12*	2		
1·6— 2·0	13 } 18	21·0	20·3	*·12—·24*	4·5		
2·0— 2·4	24	33·1	33·3	*·24—·36*	10·5 } 52	38·5	43·7
2·4— 2·8	51	55·0	55·5	*·36—·48*	15		
2·8— 3·2	72	76·1	76·7	*·48—·60*	20		
3·2— 3·6	106	91·8	92·2	*·6— 1·2*	84·5	100·5	96·7
3·6— 4·0	99	99·7	99·7	*1·2— 1·8*	111·5	117·2	114·2
4·0— 4·4	104	99·8	99·6	*1·8— 2·4*	110	114·5	113·2
4·4— 4·8	95	93·9	93·6	*2·4— 3·0*	95	103·7	103·6
4·8— 5·2	76	84·1	83·8	*3·0— 3·6*	101	90·3	90·7
5·2— 5·6	75	72·4	72·1	*3·6— 4·2*	87	76·8	77·3
5·6— 6·0	49	60·3	60·1	*4·2— 4·8*	57·5	64·2	64·7
6·0— 6·4	49	49·0	48·9	*4·8— 5·4*	49·5	53·1	53·6
6·4— 6·8	37	39·0	39·0	*5·4— 6·0*	48·5	43·7	44·1
6·8— 7·2	39	30·4	30·4	*6·0— 6·6*	38·5	35·9	36·1
7·2— 7·6	36	23·4	23·5	*6·6— 7·2*	31·5	29·2	29·4
7·6— 8·0	14	17·9	17·9	*7·2— 7·8*	30·5	23·9	24·0
8·0— 8·4	12	13·5	13·6	*7·8— 8·4*	22	19·5	19·5
8·4— 8·8	9	10·1	10·2	*8·4— 9·0*	11	15·9	15·9
8·8— 9·2	15 } 23	13·2	13·3	*9·0— 9·6*	11	13·0	13·0
9·2— 9·6	8			*9·6—10·2*	11	10·6	10·6
9·6—10·0	1			*10·2—10·8*	10 } 16	15·8	15·7
10·0—10·4	1			*10·8—11·4*	6		
10·4—10·8	5			*11·4—12·0*	7 } 10	10·6	10·5
10·8—11·2	1			*12·0—12·6*	3		
11·2—11·6	—			*12·6—13·2*	6·5		
11·6—12·0	—			*13·2—13·8*	1·5 } 11	9·9	9·8
12·0—12·4	1 } 12	16·3	16·3	*13·8—14·4*	3		
12·4—12·8	1			*14·4—15·0*	3		
12·8—13·2	—			*15·0—15·6*	3		
13·2—13·6	—			*15·6—16·2*	—		
13·6—14·0	1			*16·2—16·8*	1		
⋮	—			*16·8—17·4*	1		
17·2—17·6	1			*17·4—18·0*	— } 11	13·2	13·7
				18·0—18·6	—		
				18·6—19·2	1		
				19·2—19·8	1		
				19·8—20·4	—		
				20·4—21·0	1		
Totals	1000	1000	1000	Totals	1000	1000	1000

or the curve for the distribution of σ is

$$y = 2y_0 \frac{\sigma^{2q_2+1}}{\left\{\sigma^2 + \Sigma^2\left(1-\frac{1}{n}\right)\frac{q_1-q_2-2}{q_2+1}\right\}^{q_1}} \quad \text{.....(21).}$$

Let us see what happens to this curve as we approach the Gaussian point $\beta_1=0$, $\beta_2=3$, $\beta_4=15$. We have from (7) and (8)

$$\gamma_1=\frac{n+1}{n-1}, \quad \gamma_2=\frac{n+3}{n-1}.$$

Hence by (15) $$q_2=\tfrac{1}{2}(n-1)-1=\tfrac{1}{2}(n-3),$$

or $$2q_2+1=n-2.$$

But $2\gamma_1-\gamma_2-1=0$, and accordingly as we approach the Gaussian point $q_1\to\infty$. We can therefore write equation (21) in the form

$$y=\text{const.}\times\frac{\sigma^{n-2}}{\left(1+\dfrac{n\sigma^2}{2\Sigma^2 q_1}\right)^{q_1}},$$

where $q_1\to\infty$, i.e.

$$y=y_0'\sigma^{n-2}e^{-\frac{n\sigma^2}{2\Sigma^2}},$$

but this is the well-known distribution curve of σ for sampling from a normal population. This gives us confidence in equation (21).

The distributions obtained by sampling are given in Table IV (p. 409), which shows also the frequencies obtained by using equation (21). All the constants for the equations (y_0, q_1, q_2, a, origin and start of curve) have already been given when dealing with the variance. The mid-ordinates of each group were computed and corrected to frequency by using the equation

Frequency = Mid-ordinate

$$\times h\left\{1+\frac{2h^2}{24}\left[\frac{x^4(2q_2-2q_1+1)(q_2-q_1)+x^2a(4q_2{}^2-4q_1q_2+2q_2-3q_1)+a^2q_2(2q_2+1)}{x^2(x^2+a)^2}\right]\right\}.$$

For the Mode $$x=\sqrt{\frac{a(2q_2+1)}{2q_1-2q_2-1}}.$$

The degrees of goodness of fit of these curves agree closely with those obtained for the distribution of the variance. The curves are shown in Diagrams VII and VIII (see pp. 410—11).

7. *Distribution of "Student's"* $z=\dfrac{m-\overline{m}}{\sigma}$ *in Samples.*

The distributions and frequency constants of our samples are given in Table V, p. 412. If the samples had been drawn from a normal population, the distributions of z would be symmetrical, and would be fitted from "Student's" tables in *Metron*, Vol. v. (1925) where $t=z\sqrt{n'-1}$. Our distributions are however sufficiently skew to give a poor fit with samples of 20 and a still worse fit with samples of 5. The distributions of z show then that the samples could not have been drawn from a normal population. The positive skewness of the population has resulted in negative skewness in the distributions of z in samples. A similar result was found by Neyman and Pearson using the z distribution from one of Church's sets of samples*, when negative skewness in the population gave positive skewness in the distribution of z in samples of 10.

* *Biometrika*, Vol. xxa. p. 198.

TABLE IV.

Distribution of Standard Deviation of Samples.

$n=20$			$n=5$		
Value of σ	Observed	Equation (21)	Value of σ	Observed*	Equation (21)
·47— ·51	1		·11— ·19	3	
·51— ·55	1	19·4	·19— ·27	14	17·0
·55— ·59	7		·27— ·35	37	27·8
·59— ·63	9		·35— ·43	45	46·9
·63— ·67	16	18·6	·43— ·51	56	66·8
·67— ·71	18	29·8	·51— ·59	83	84·1
·71— ·75	43	43·3	·59— ·67	90	96·0
·75— ·79	48	57·9	·67— ·75	94	101·1
·79— ·83	73	71·6	·75— ·83	108	99·5
·83— ·87	100	82·7	·83— ·91	105	92·4
·87— ·91	89	89·5	·91— ·99	72	81·5
·91— ·95	91	91·4	·99—1·07	68	68·8
·95— ·99	89	88·6	1·07—1·15	60	55·9
·99—1·03	83	81·7	1·15—1·23	51	43·9
1·03—1·07	71	72·1	1·23—1·31	35	33·5
1·07—1·11	47	61·1	1·31—1·39	20	24·9
1·11—1·15	57	49·8	1·39—1·47	21	18·1
1·15—1·19	40	39·3	1·47—1·55	13	12·9
1·19—1·23	37	30·0	1·55—1·63	10	15·3
1·23—1·27	28	22·3	1·63—1·71	5	
1·27—1·31	10	16·1	1·71—1·79	5	
1·31—1·35	19	11·3	1·79—1·87	2	
1·35—1·39	12	13·2	1·87—1·95	1	13·6
1·39—1·43	1		1·95—2·03	1	
1·43—1·47	5		2·03—2·11	1	
1·47—1·51	1				
1·51—1·55	—				
1·55—1·59	2				
1·59—1·63	—				
1·63—1·67	1	10·3			
1·67—1·71	—				
1·71—1·75	—				
1·75—1·79	—				
1·79—1·83	—				
1·83—1·87	1				
Totals	1000	1000	Totals	1000	1000
Mean ... ·96796		Mode = ·9252	Mean ... ·8346		Mode = ·7287
σ ... ·17797		—	σ ... ·32285		—
β_1 ... ·2297		$\chi^2 = 25·489$	β_1 ... ·2625		$\chi^2 = 13·219$
β_2 ... 3·5650		$P =$ ·18	β_2 ... 3·1814		$P =$ ·78
Skewness +0·2112		—	Skewness +0·2970		—

* N.B. Twice the above numbers of groups were used to find the frequency constants.

Best fitting curves for each of our distributions were of Type IV and were fitted by computing ordinates, drawing the curves and using the planimeter to find

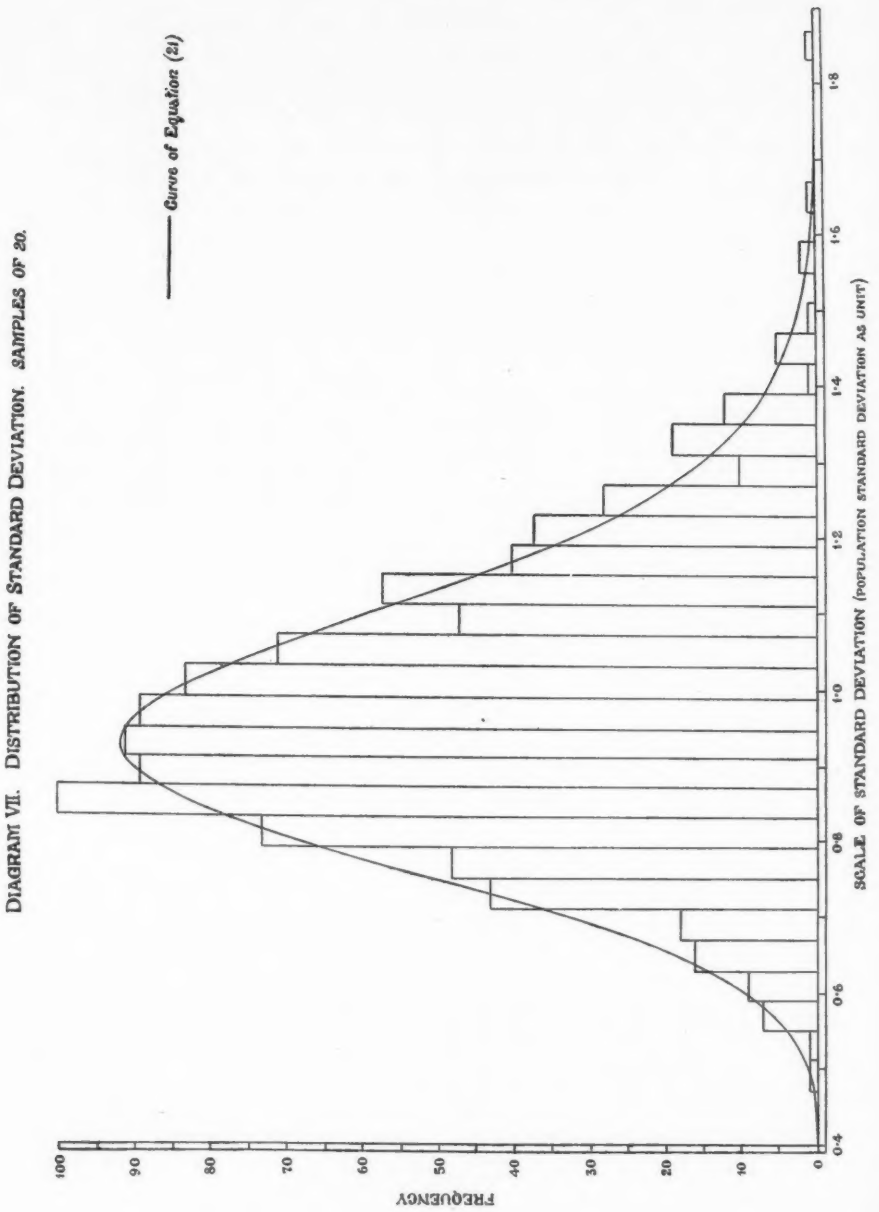


DIAGRAM VII. DISTRIBUTION OF STANDARD DEVIATION. SAMPLES OF 20.

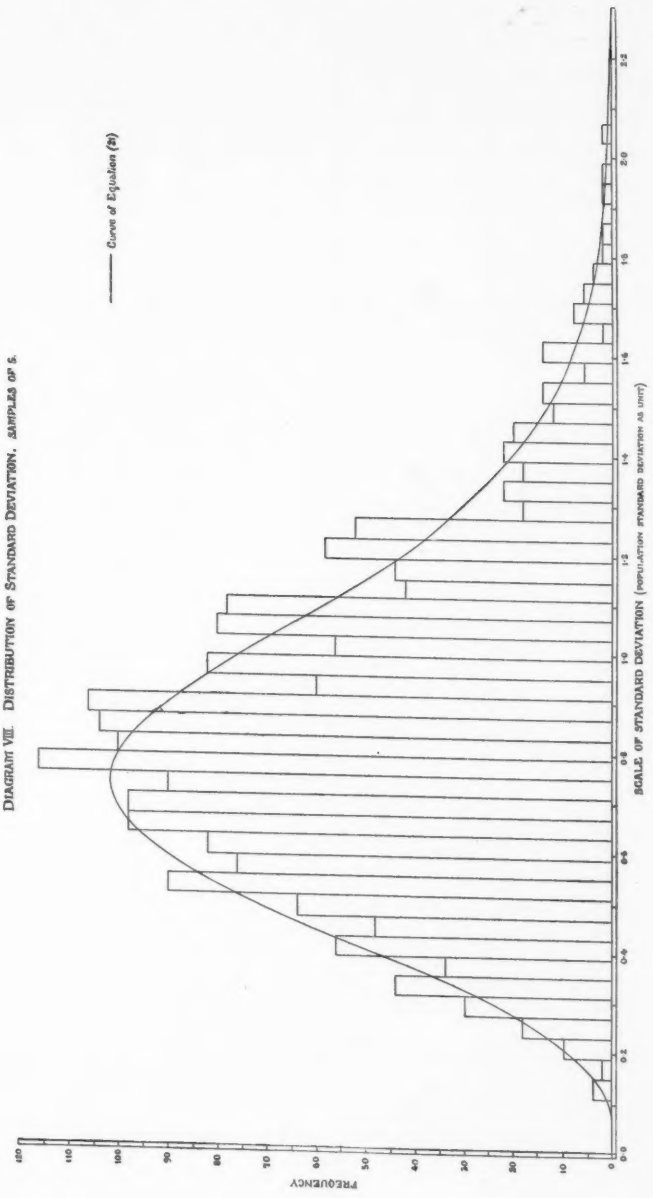


DIAGRAM VIII. DISTRIBUTION OF STANDARD DEVIATION. SAMPLES OF 5.

TABLE V. $z = \dfrac{m - \overline{m}}{\sigma}$ *in Samples.*

$n=20$					$n=5$				
Values of z	Observed		Best Fitting Curve	Normal Theory	Equation (26)[b]	Values of z	Observed		Best Fitting Curve	Normal Theory	Equation (26)[b]
$-\cdot95$ — $-\cdot90$	2					$-5\cdot9$ — $-5\cdot8$	1				
$-\cdot90$ — $-\cdot85$	—					⋮					
$-\cdot85$ — $-\cdot80$	3	11	11·5	5·3	10·2	$-3\cdot8$ — $-3\cdot7$	1				
$-\cdot80$ — $-\cdot75$	3					$-3\cdot7$ — $-3\cdot6$	1				
$-\cdot75$ — $-\cdot70$	3					⋮					
$-\cdot70$ — $-\cdot65$	—					$-3\cdot1$ — $-3\cdot0$	2				
$-\cdot65$ — $-\cdot60$	6	14	13·3	8·2	11·6	$-3\cdot0$ — $-2\cdot9$	1	15	15·1	6·8	12·2
$-\cdot60$ — $-\cdot55$	8					⋮					
$-\cdot55$ — $-\cdot50$	12		10·6	7·8	9·9	$-2\cdot6$ — $-2\cdot5$	1				
$-\cdot50$ — $-\cdot45$	15		14·6	11·4	13·8	$-2\cdot5$ — $-2\cdot4$	2				
$-\cdot45$ — $-\cdot40$	24		19·8	16·5	19·0	$-2\cdot4$ — $-2\cdot3$	2				
$-\cdot40$ — $-\cdot35$	21		26·3	23·1	25·7	$-2\cdot3$ — $-2\cdot2$	2				
$-\cdot35$ — $-\cdot30$	29		33·5	31·2	33·0	$-2\cdot2$ — $-2\cdot1$	2				
$-\cdot30$ — $-\cdot25$	51		42·3	41·5	42·5	$-2\cdot1$ — $-2\cdot0$	3				
$-\cdot25$ — $-\cdot20$	39		51·9	52·8	52·7	$-2\cdot0$ — $-1\cdot9$	1	11	13·0	6·8	10·6
$-\cdot20$ — $-\cdot15$	59		61·6	63·5	63·3	$-1\cdot9$ — $-1\cdot8$	4				
$-\cdot15$ — $-\cdot10$	78		70·4	73·2	72·4	$-1\cdot8$ — $-1\cdot7$	3				
$-\cdot10$ — $-\cdot05$	79		77·5	80·6	80·2	$-1\cdot7$ — $-1\cdot6$	3	12	11·5	6·4	9·0
$-\cdot05$ — 0	78		81·9	84·9	85·5	$-1\cdot6$ — $-1\cdot5$	9				
0 — $\cdot05$	96		83·0	84·9	86·3	$-1\cdot5$ — $-1\cdot4$	3	14	16·6	10·0	14·0
$\cdot05$ — $\cdot10$	82		80·0	80·6	82·1	$-1\cdot4$ — $-1\cdot3$	11				
$\cdot10$ — $\cdot15$	65		73·2	73·2	74·9	$-1\cdot3$ — $-1\cdot2$	15		11·2	7·2	9·8
$\cdot15$ — $\cdot20$	66		64·4	63·5	64·3	$-1\cdot2$ — $-1\cdot1$	13		13·2	9·1	12·0
$\cdot20$ — $\cdot25$	58		53·6	52·8	52·1	$-1\cdot1$ — $-1\cdot0$	14		16·2	11·8	14·6
$\cdot25$ — $\cdot30$	39		41·0	41·5	39·9	$-1\cdot0$ — $\cdot9$	15		19·2	15·0	18·8
$\cdot30$ — $\cdot35$	32		30·3	31·2	29·0	$-\cdot9$ — $\cdot8$	22		23·2	19·3	23·4
$\cdot35$ — $\cdot40$	14		21·9	23·1	20·0	$-\cdot8$ — $\cdot7$	35		27·2	24·7	28·8
$\cdot40$ — $\cdot45$	15	23	23·8	27·9	20·9	$-\cdot7$ — $\cdot6$	37		33·0	31·1	35·6
$\cdot45$ — $\cdot50$	8					$-\cdot6$ — $\cdot5$	34		38·6	38·8	43·2
$\cdot50$ — $\cdot55$	7					$-\cdot5$ — $\cdot4$	34		44·8	47·3	51·6
$\cdot55$ — $\cdot60$	4					$-\cdot4$ — $\cdot3$	53		51·2	56·1	60·4
$\cdot60$ — $\cdot65$	1	15	13·6	21·3	10·7	$-\cdot3$ — $\cdot2$	70		57·6	64·4	68·0
$\cdot65$ — $\cdot70$	2					$-\cdot2$ — $\cdot1$	65		63·6	70·8	73·8
$\cdot70$ — $\cdot75$	1					$-\cdot1$ — 0	66		68·2	74·4	77·0
						0 — $\cdot1$	68		70·6	74·4	74·6
						1 — $\cdot2$	86		70·4	70·8	70·0
						$\cdot2$ — $\cdot3$	62		67·2	64·4	62·2
						$\cdot3$ — $\cdot4$	67		61·6	56·1	52·8
						$\cdot4$ — $\cdot5$	63		53·6	47·3	43·4
						$\cdot5$ — $\cdot6$	32		44·0	38·8	32·8
						$\cdot6$ — $\cdot7$	29		34·2	31·1	25·2
						$\cdot7$ — $\cdot8$	23		25·0	24·7	19·4
						$\cdot8$ — $\cdot9$	11		17·8	19·3	14·4
						$\cdot9$ — 1·0	7	13	19·6	26·8	18·6
						1·0 — 1·1	6				
						1·1 — 1·2	6				
						1·2 — 1·3	4				
						1·3 — 1·4	6				
						1·4 — 1·5	1				
						1·5 — 1·6	—	21	12·6	46·3	23·8
						1·6 — 1·7	1				
						1·7 — 1·8	1				
						1·8 — 1·9	1				
						⋮					
						2·2 — 2·3	1				
Totals	1000		1000	1000	1000	Totals	1000		1000	1000	1000
χ^2			15·578	33·214	17·607	χ^2			32·728	77·07	39·307
P			·835	·049	·728	P			·266	$<$·000,001	·062

Frequency Constants.

	$n=20$	$n=5$
Mean ...	$-\cdot02235 \pm \cdot00532$	$-\cdot1283 \pm \cdot0149$
σ ...	$\cdot24938 \pm \cdot00376$	$\cdot7005 \pm \cdot0106$
β_1 ...	$\cdot1059 \pm \cdot0448$	1·9441
β_2 ...	$3\cdot4386 \pm \cdot2415$	9·5287
Skewness ...	$-\cdot1386$	$-\cdot3238$
k ...	·1460	·3271
Type ...	IV	IV
Mode ...	$+\cdot0123$	$+\cdot0985$

frequencies. They are shown in Diagrams IX and X. The constants of these are as follows, if the equation to the curve be:

$$y = y_0 e^{-\nu \tan^{-1}\frac{x}{a}} \left(1+\frac{x^2}{a^2}\right)^{-m},$$

	$n=20$	$n=5$
y_0 ...	240·5140	278·564
ν ...	10·342,806	3·812,3246
a ...	1·129,3891	1·214,577
m ...	13·507,681	3·734,053
Origin ...	+0·4447 } from	+0·7185 } from
Mode ...	+0·0123 } $z=0$	+0·0985 } $z=0$

The frequencies for these curves and those for the normal theory are shown in Table V (p. 412), where also are given the P values for goodness of fit.

Professor Karl Pearson has made the following attempt to provide a general equation for the distribution of z in samples, using the above results and those of Church's sampling.

8. *An approximate Method of finding the Distribution of z in Samples.*

The results of sampling for 5, 10 and 20 series from skew populations lead us to anticipate a Type IV curve for the distribution of z, i.e. of the form

$$y = y_0 \frac{e^{-\nu \tan^{-1}\frac{x}{a}}}{\left(1+\frac{x^2}{a^2}\right)^m} \quad \text{.....(22),}$$

where the relation of x to z has to be determined.

If the skewness be zero and we sample from a symmetrical population we have frequencies nearly satisfied by "Student's" curve, for which we may take $x=z$,

$$y = \frac{y_0}{(1+z^2)^{\frac{n}{2}}} \quad \text{.....(23),}$$

if we are not too far removed from the normal point $\beta_1=0$, $\beta_2=3$, along the axis $\beta_1=0$.

It seems therefore reasonable to suppose that m, a and ν only differ from $\frac{1}{2}n$, 1 and 0 by small quantities depending on the skewness of the sampled population. Further if capitals denote the values in the original population and small letters values in the population of samples, then for distribution (23) of z

$$\sigma^2 = \frac{1}{n-3} \quad \text{.....(24).}$$

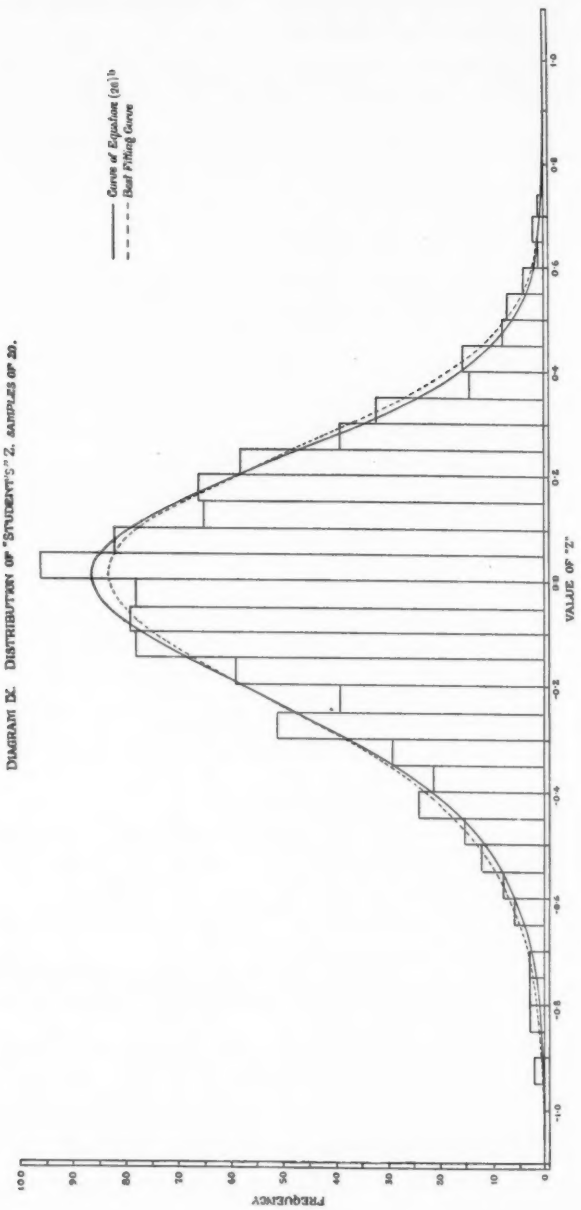


DIAGRAM IX. DISTRIBUTION OF "STUDENT'S" Z, SAMPLES OF 20.
Curve of Equation (28)b
Best Fitting Curve
VALUE OF "Z"
FREQUENCY

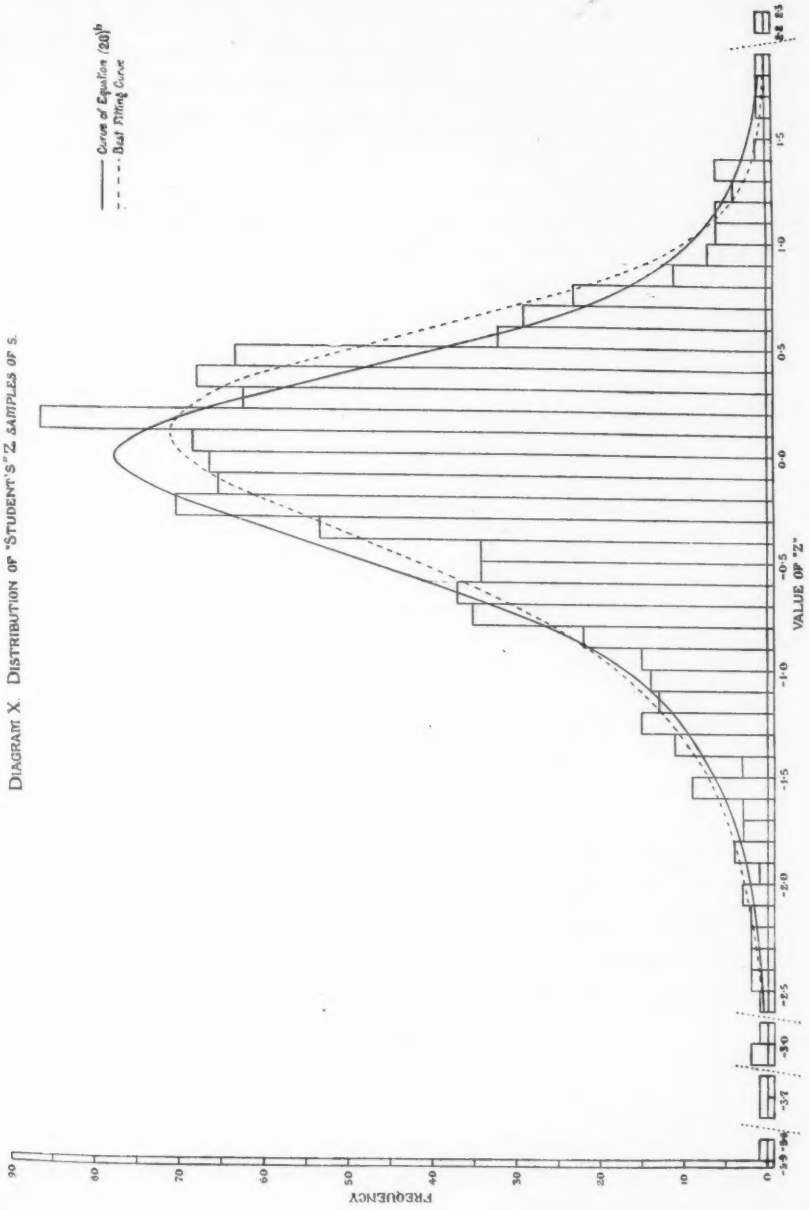


DIAGRAM X. DISTRIBUTION OF "STUDENT'S" Z, SAMPLES OF 5.

Now the samplings chosen lie fairly away from the normal point, namely:

	$n=5$	$n=10$ (Church)*	$n=20$
B_1 ...	0·497,025	0·219,333	0·497,025
B_2 ...	3·729,789	3·157,676	3·729,789
Σ† ...	0·99986	1·00090	0·99986

Now let us suppose our m in (22) to be one half of n, the size of the samples. Then we have, if β_1 and β_2 be the β's of the distribution of samples,

$$n = \nu + 2 = \frac{6(\beta_2 - \beta_1 - 1)}{2\beta_2 - 3\beta_1 - 6} + 2 = \frac{10\beta_2 - 12\beta_1 - 18}{2\beta_2 - 3\beta_1 - 6},$$

or

$$\beta_2 = \frac{(3n-12)\beta_1 + 6(n-3)}{2(n-5)} \quad \text{.....(25).}$$

Now the observed values for the three samples were:

	$n=5$	$n=10$	$n=20$
σ ...	·7005	·3709	·2494
β_1 ...	1·9441	·1740	·1059
β_2 ...	9·5287	4·5405	3·4386

while calculated from (24), (25) and the observed β_1 they are:

	$n=5$	$n=10$	$n=20$
σ ...	·7071	·3780	·2425
β_2 ...	∞	4·5132	3·5694

We have found that taking $\beta_2 = \infty$, when $n = 5$, gives a quite reasonable fit to the distribution of z samples, although its observed β_2 is only 9·5287; we conclude accordingly that our values for σ and β_2 as found from (24) and (25) will likewise give a reasonable fit in the other two cases, as they are not more divergent from observed values.

If now we assume the form of the curve of frequency of z for samples to be

$$y = y_0 \frac{e^{-\nu \tan^{-1}\frac{x}{a}}}{\left(1 + \frac{x^2}{a^2}\right)^{\frac{1}{2}n}} \quad \text{.....(26),}$$

* Values calculated by Neyman and Pearson, *Biometrika*, Vol. XXA. p. 198.

† Taken as unity for the continuous population curve. The above values result from grouping for sampling.

we have
$$\nu^* = \pm \frac{(n-2)\sqrt{\frac{1}{16}\beta_1\frac{(n-4)^2}{n-3}}}{\sqrt{1-\frac{1}{16}\beta_1\frac{(n-4)^2}{n-3}}} \quad \ldots\ldots(27),$$

$$a = \sqrt{\frac{\sigma^2(n-3)}{(n-2)^2}\left\{1-\frac{1}{16}\beta_1\frac{(n-4)^2}{n-3}\right\}} \times (n-2)$$

$$= \sqrt{1-\frac{1}{16}\beta_1\frac{(n-4)^2}{n-3}} \quad \ldots\ldots(28),$$

since $\sigma^2 = \frac{1}{n-3}$ approximately.

Further the mean value of x, or $\bar{x}$ is given by

$$\bar{x} = \frac{-a\nu}{(n-2)}$$

$$= +\tfrac{1}{4}\sqrt{\beta_1}\frac{(n-4)}{\sqrt{n-3}} \quad \ldots\ldots(29).$$

Distance from Origin to Mode

$$= -\frac{a\nu}{n}$$

$$= +\tfrac{1}{4}\sqrt{\beta_1}\frac{(n-4)}{\sqrt{n-3}}\frac{(n-2)}{n} \quad \ldots\ldots(30).$$

Distance from Mean to Mode or d

$$= +\tfrac{1}{2}\sqrt{\beta_1}\frac{n-4}{n\sqrt{n-3}} \quad \ldots\ldots(31).$$

We have next to consider what is the appropriate relation between x and z. It by no means follows that we shall get the best fitting curve (26) by taking $x = z$. If we do, we shall get indeed the simplest form of curve, for the origin of x will be the origin of z or $z = 0$, and our equations reduce to

$$\nu = \frac{-(n-2)\bar{z}}{\sqrt{1-\bar{z}^2}} \quad \ldots\ldots(27)^a,$$

$$a = \sqrt{1-\bar{z}^2} \quad \ldots\ldots(28)^a,$$

$$\bar{z} = +\tfrac{1}{4}\sqrt{\beta_1}\frac{(n-4)}{\sqrt{n-3}} \quad \ldots\ldots(29)^a,$$

where $\bar{z}$ is to be given the same sign as $\sqrt{\beta_1}$, i.e. that of the third moment of the population of samples, and

$$y = y_0 \frac{e^{+\frac{(n-2)\bar{z}}{\sqrt{1-\bar{z}^2}}\tan^{-1}\frac{z}{\sqrt{1-\bar{z}^2}}}}{\left(1+\frac{z^2}{1-\bar{z}^2}\right)^{\frac{n}{2}}} \quad \ldots\ldots(26)^a.$$

* The sign to be given to ν is always opposite to that of μ_3, or of $\sqrt{\beta_1}$, i.e. if $\sqrt{\beta_1}$ be positive, ν is to be taken negative, for

$$\mu_3 = -\frac{4\nu}{(n-3)(n-4)}\sqrt{1-\frac{1}{16}\beta_1\frac{(n-4)^2}{n-3}}.$$

Phil. Trans. Vol. 186A, 1895, pp. 377–8.

Since $\bar{z}$ is in our sampling experiments a small quantity we might even take this last as

$$y = y_0 \frac{e^{+(n-2)\bar{z}\tan^{-1}z}}{(1+z^2)^{\frac{1}{2}n}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (26)^{a'},$$

taking $\qquad v = -(n-2)\bar{z}$ and $a = 1$.

It is quite possible that $(26)^{a'}$ would give for much sampling work a reasonable description of the effect of skewness of the sampled population on the population of samples. But have we actual evidence to show that we are justified in taking the x of (26) to be identical with z? We might naturally expect that they would differ by a quantity of the order $\sqrt{\beta_1}$. Let us see what happens in the case of our three samples if we use $(29)^a$.

	$n=5$	$n=10$	$n=20$
$\bar{z}$	− ·1283	+ ·0120	− ·02235
$\sqrt{\beta_1}$ ...	− 1·394,317	+ ·417,114	− ·325,424
$\bar{x}$ by (29) ...	− ·2465	+ ·2365	− ·3157

It is clear that $\bar{x}$ does not approach even approximately to $\bar{z}$; thus we cannot suppose the origin of our curve (26) to be at $z=0$, i.e. x and z differ very substantially.

The equation corresponding to (26) using (29) is accordingly:

$$y = y_0 \frac{e^{+(n-2)\frac{\bar{x}}{\sqrt{1-\bar{x}^2}}\tan^{-1}\frac{x}{\sqrt{1-\bar{x}^2}}}}{\left(1+\frac{x^2}{1-\bar{x}^2}\right)^{\frac{1}{2}n}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (26)^{b},$$

where $\qquad \bar{x} = +\tfrac{1}{4}\sqrt{\beta_1}\,\dfrac{n-4}{\sqrt{n-3}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (29)^{b}.$

Thus it follows that we need two constants of our population of samples of z, i.e. $\bar{z}$, the mean z, in order to place $(29)^b$ down on the observations, and $\sqrt{\beta_1}$ in order to obtain the origin of the curve $(29)^b$. In other words we must either take $\bar{z}$ and $\sqrt{\beta_1}$ from our population of samples—and these will not be very reliable if we have only a small population of samples—or we must determine the first three moment coefficients of z from a knowledge of the relations of m and Σ in the skew population sampled, i.e. we must find $\bar{z}$ and $\sqrt{\beta_1}$ in terms of the constants of the sampled population. This latter problem has been tried from many aspects and found to be of extraordinary difficulty, arising from the non-convergency of the series which must be used for expanding $1/\sigma$. For finite range populations—and actually in practice all curves are of this nature—the theoretical distribution of z will be like those of m and σ, i.e. will consist of a number of separate curves, so that we could only hope for representation by a single curve as n grows larger, as

an approximation. A better result should probably be obtained in curves of infinite range, when a single curve for the distribution of z is reasonably to be expected.

A final attempt to obtain the moments of z in samples from a skew population was made on the following assumptions:

(1) That the mode of z must be close to the mode of the curve of means.

(2) That to determine the modal value of z we may give to σ the value which is by correlation the most probable value of σ for the given m.

Let B_1 and B_2 be the β's of the sampled population: then the β's of the curve of means are for samples of size n,

$$\beta_1 = \frac{B_1}{n}, \quad \beta_2 = 3 + \frac{(B_2 - 3)}{n},$$

and consequently (on the assumption that a Pearson curve fits the distribution) the distance of the mode from the mean

$$= \frac{-\sqrt{\beta_1}(\beta_2 + 3)}{2(5\beta_2 - 6\beta_1 - 9)} \frac{\Sigma}{\sqrt{n}},$$

$$= \frac{-\sqrt{\frac{B_1}{n}}\left(6 + \frac{B_2 - 3}{n}\right)}{2\left\{6 + \frac{5(B_2 - 3)}{n} - 6\frac{B_1}{n}\right\}} \frac{\Sigma}{\sqrt{n}},$$

since, if Σ be the standard deviation of the sampled population, $\Sigma/\sqrt{n}$ is the corresponding standard deviation of the curve of means. Let $\tilde{\mu}_2$ denote the value of μ_2 in samples which correspond to the modal m on the regression line of μ_2 on m, then

$$\tilde{\mu}_2 - \bar{\mu}_2 = \frac{r_{m\mu_2}\sigma_{\mu_2}}{\sigma_m}(\tilde{m} - \bar{m}).$$

But* $$\bar{\mu}_2 = \frac{n-1}{n}\Sigma^2, \quad \sigma_m = \frac{\Sigma}{\sqrt{n}},$$

$$\sigma_{\mu_2} = \frac{n-1}{n}\frac{\Sigma^2}{\sqrt{n}}\sqrt{B_2 - \frac{n-3}{n-1}},$$

$$r_{m\mu_2} = \frac{\sqrt{B_1}}{\sqrt{B_2 - \frac{n-3}{n-1}}}.$$

Consequently $$\tilde{\mu}_2 = \frac{n-1}{n}\Sigma^2 + \frac{\frac{n-1}{n}\frac{\Sigma^2}{\sqrt{n}}\sqrt{B_1}(\tilde{m} - \bar{m})}{\frac{\Sigma}{\sqrt{n}}}$$

$$= \frac{n-1}{n}\frac{\Sigma^2}{2}\left\{1 - \frac{\frac{B_1}{n}\left(6 + \frac{B_2 - 3}{n}\right)}{6 + \frac{5(B_2 - 3)}{n} - 6\frac{B_1}{n}}\right\},$$

* *Biometrika*, Vol. XVII. p. 479.

and
$$\tilde{z} = \frac{-\sqrt{\frac{B_1}{n(n-1)}}\,\frac{6+\frac{B_2-3}{n}}{2\left\{6+\frac{5(B_2-3)}{n}-6\frac{B_1}{n}\right\}}}{\sqrt{1-\frac{\frac{B_1}{2n}\left(6+\frac{B_2-3}{n}\right)}{\left(6+\frac{5(B_2-3)}{n}-6\frac{B_1}{n}\right)}}}.$$

Hence, if we compute
$$\lambda = \tfrac{1}{2}\frac{\left(6+\frac{B_2-3}{n}\right)}{\left(6+\frac{5(B_2-3)}{n}-6\frac{B_1}{n}\right)},$$

we shall have
$$\tilde{z} = \frac{-\lambda\sqrt{\frac{B_1}{n(n-1)}}}{\sqrt{1-\frac{B_1}{n}\lambda}}.$$

These values were computed and compared with the modal values obtained from the best fitting curves of our samples of 5 and 20, and also for Church's samples of 10.

Modal Values of z.

	Computed	Observed
$n=5$	$-\cdot0810$	$+\cdot0985$
$n=10$	$+\cdot0252$	$-\cdot0341$
$n=20$	$-\cdot0182$	$+\cdot0123$

These results, though not numerically widely out, are all of *opposite* sign to the observed, and show that either our hypothesis or our algebra is incorrect. Checking showed no error. We are accordingly forced back to equations (26)^b and (29)^b as being the best we can find at present to give a reasonable description of our sampling experiments.

These equations were tested by fitting the curves to our distributions of z, and to Church's samples of 10. The constants are given below, and frequencies obtained by drawing the curves and using the planimeter are given in Table V (p. 412). Curves are shown in Diagrams IX and X (pp. 414–15).

$$y = y_0\,\frac{e^{+(n-2)\frac{\bar{x}}{\sqrt{1-\bar{x}^2}}\tan^{-1}\frac{x}{\sqrt{1-\bar{x}^2}}}}{\left(1+\frac{x^2}{1-\bar{x}^2}\right)^{\frac{1}{2}n}},$$

where
$$\bar{x} = +\tfrac{1}{4}\sqrt{\beta_1}\,\frac{n-4}{\sqrt{n-3}}.$$

As shown at the foot of Table V, the fit for samples of 20 is very satisfactory, while that for samples of 5 is not bad when compared to the fit of the best fitting curve. For Church's samples of 10, the best fitting curve gave $\chi^2 = 20{\cdot}875$ and $P = {\cdot}23$ for 18 groups, and the curve from equation (26)^b $\chi^2 = 20{\cdot}903$ and $P = {\cdot}23$ for 18 groups, which is also satisfactory.

	$n=20$	$n=10$	$n=5$
Mean ...	−·02235	+·0120	−·1283
Mode ...	+·0093	−·0353	−·0297
Origin ...	+·2934	−·2245	+·1182
y_0 ...	715·0182	972·9007	720·8537
ν ...	5·989,0381	−1·947,078	·762,9886
a ...	·948,8565	·971,636	·969,1472
m ...	10	5	2·5

9. *The Use of "Student's" Tables when the Population is skew.*

When the above samples of z had been obtained, we were able to consult "Student" and discuss the results. He has pointed out that in practice it often happens that a single small sample supplies all the information which we have about a population which we may have reason to assume is skew, but we may not know the direction of skewness. We wish to know if a given value z_p is too high to be likely to occur by chance, and it is necessary to consider, not the chance of z being greater than z_p on the supposition that the original distribution is positively skew or on the supposition that it is negatively skew, but, since either is equally likely, one half of the sum of these chances.

This is equivalent to working from a new symmetrical distribution of z obtained by replacing the number of z's by half the sum of the z's positive and negative at each value.

To test this, we have taken our distributions of z in samples and "folded them over" the line $z = 0$. For example, the frequencies in our samples of 20 from the group 0 to ·05 were added to the frequencies from the group 0 to −·05 and the sum compared to twice the frequency for that group from "Student's" Tables for $n = 20$ and so on. The results were as follows:

$$n = 20, \text{ for 13 groups } \chi^2 = 7{\cdot}58, \quad P = {\cdot}816;$$

$$n = 5, \text{ for 17 groups } \chi^2 = 14{\cdot}81, \quad P = {\cdot}539.$$

We find then, that the new distribution is not very different from that which would be obtained in sampling from a normal population, or more probably from some symmetrical population.

So if we use "Student's" Tables to decide whether a given sample has been drawn from a population of which we only know the mean, these results suggest that we

shall be right *in the long run* provided that the skewness of the population does not exceed that used in this paper*.

10. *Correlations.*

The following correlation coefficients were calculated from our two sets of samples:

	$n=20$	$n=5$
Means and Variances ... r	$+\cdot4158\pm\cdot0176$	$+\cdot3996\pm\cdot0179$
" " " ... η	$\cdot4488$	$\cdot4416$
Means and Standard Deviations r	$+\cdot4242\pm\cdot0175$	$+\cdot4059\pm\cdot0178$
" " " ... η	$\cdot4458$	$\cdot4469$

Neyman† has found that the correlation of means and variances in samples drawn from an infinite population is given by

$$r=\sqrt{\frac{\beta_1}{\beta_2-\frac{n-3}{n-1}}}.$$

Values calculated from the β's of our population are $+\cdot4187$ for $n=20$ and $+\cdot3923$ for $n=5$, which agree very well with the values observed. One would expect the correlation between the means and standard deviations to be almost exactly the same as that between the means and variances owing to the high correlation between variance and standard deviation as shown by Yasukawa's work‡, but no doubt the differences shown above are due to grouping of the data.

It is interesting to note how closely the four correlation ratios agree among themselves.

11. *Summary.*

Two sets of random samples have been drawn from an infinite population of definite skewness and of Type III: 1000 samples of 5 and 1000 samples of 20.

The distributions of the means of samples were adequately described by normal curves, but with the smaller samples a slight improvement in fit was found with the theoretical β's which give a Type III curve.

The distributions of ranges were well fitted in the case of the smaller samples by the normal theory due to Tippett, but less well in the case of the larger samples. Best fitting curves (of similar type in each case to the normal theory) are shown.

The distributions of variances were adequately described by Type VI curves based on the moments calculated by Church's formula from the higher β's up to β_6 of the population. Professor Pearson has devised an equation for the distribution of variances, assuming a curve of Type VI which starts at zero and has its origin at

* [Supposing 50 % of prisoners tried for murder were acquitted and the remainder found guilty, should we be right *in the long run* to drop the trial and toss up for judgment? Ed.]

† *Biometrika*, Vol. XVII. p. 479.

‡ *Biometrika*, Vol. XVII. p. 213.

zero, thereby using the constants of the population distribution only up to β_4. Curves of this type give satisfactory fits, as also did curves for the distributions of standard deviations using the same constants.

The distributions of "Student's" z were too markedly skew to be fitted by the normal theory, particularly in the case of the smaller samples. They were fitted with curves of Type IV. Professor Pearson has attempted to devise an equation for the distribution of z, assuming Type IV, which will reduce to "Student's" curve for samples drawn from a normal population. He was unable to do so using the constants of the sampled population, but devised an equation for which we must know the mean and $\sqrt{\beta_1}$ of z in the population of samples. This curve gave satisfactory fits with both sets of samples and also with a set of samples used by Church.

The value of "Student's" Table in practice, even when the population is definitely skew, is indicated.

The correlation coefficients between the means and variances, means and standard deviations are given, and they differ very sensibly from zero, the value in the case of a normal population.

I should like to express my very great thanks to Professor Pearson for the assistance he has so freely given during this work and for the privilege of working under him at University College. Also to Dr E. S. Pearson for constant help and advice, and to "Student," on whose initiative the work was begun and who has given valuable help and criticism. My thanks are due also to "Mathetes" who drew the diagrams so well.

MISCELLANEA.

I. Note on the Paper by Dr J. Wishart in the present Volume (XX^A).

(See pp. 32–52.)

On p. 42 of the current volume (xx^A) of *Biometrika*, the words "of the sample" were introduced in error in line 4, and as this may tend to obscure the exact meaning of the moment coefficients listed on the succeeding pages, a word of explanation is necessary. From an infinite normal n variate population a sample of size N is taken and the n variances and $\frac{1}{2}n(n-1)$ 1st order product moments a_{pq} calculated. These values are then regarded as $\frac{1}{2}n(n+1)$ new variates, varying from sample to sample in a multiple distribution whose moment coefficients are tabulated on pp. 44—52. As calculated by the formulae (15) and (16) of p. 41, these moments were about zero, but they were subsequently corrected in the usual way to be about the *mean of the distribution* of the a's. Thus to use the 3 variate notation of p. 41, we have

$$\mu\begin{pmatrix} j & q & p \\ k & m & \\ l & & \end{pmatrix} = \int_v (a-\bar{a})^j (b-\bar{b})^k (c-\bar{c})^l (f-\bar{f})^m (g-\bar{g})^p (h-\bar{h})^q \,.\, F(a,b,c,f,g,h)\,.\,dv,$$

where v denotes multiple integration throughout the volume to be integrated and the bars denote mean values of the quantities in all possible samples.

An alternative method of calculating these moment coefficients will be described in a forthcoming number of the *Proceedings of the London Mathematical Society* (2), 29 (1929).

J. WISHART.

II. The Seven Kings of Rome.

Note by CHARLES F. TRUSTAM. (See *Biometrika*, Vol. xx^A, pp. 252—262.)

Mr W. Palin Elderton has indicated an alternative to my own actuarial method of treating this problem. I have at his suggestion worked out numerical results, which are given below. The formulae used are:

$$ {}_sk_a = l_{a+s}/l_a,$$

$$ {}_sk_{ab} = \frac{1}{l_a}\int_0^s (l_{a+t}\mu_{a+t} \times {}_{s-t}k_b)\,dt + {}_sk_a,$$

$$ {}_sk_{abc} = \frac{1}{l_a}\int_0^s (l_{a+t}\mu_{a+t} \times {}_{s-t}k_{b-c})\,dt + {}_sk_a,$$

..... etc.

In these formulae ${}_sk_a$ is the probability that a man aged a will live at least another s years, ${}_sk_{ab}$ is the similar probability in respect of two lives a and b and so on.

According to the H^M Mortality Table, the average age between 30 and 60 is $43\frac{1}{2}$. Consequently we may substitute for seven lives chosen at random between 30 and 60 the following approximate distribution: a, b each $33\frac{1}{2}$; c, d, e each $43\frac{1}{2}$; f, g each $53\frac{1}{2}$. With these values, values of ${}_sk_a$ were calculated. A series of values ${}_sk_{ab}$ were next found, the integrals being evaluated approximately by Simpson's Rule. Hence a series of values of ${}_sk_{abc}$ were computed, and so on up to ${}_sk_{abcdefg}$.

The final results were:

s	250	255	260
${}_sk_{abcdefg}$	·00598	·00350	·00183

Whence by backward differences from ·00183 we find for 257 the value

$${}_sk_{abcdefg} = ·00278.$$

By my previous work* I found the value ·00274.

It seems quite conclusive accordingly that the probability in question is of the order ·003.

* Present issue of *Biometrika* (xx^A).

III. Interpolation. By J. F. STEFFENSON, Sc.D., Professor of Actuarial Science at the University of Copenhagen. pp. ix + 248. London, Ballière, Tindall and Cox. 1927.

To English readers probably the best general description of Professor Steffensen's book is that it is an improved, rewritten and up-to-date "Boole," and the compliment we intend will be understood by anyone who remembers how good a book was Boole's as compared with any previous or contemporary work on Finite Differences.

The book is written from the mathematical point of view and necessarily calls for considerable concentration on the part of the reader but the author does not forget that the subject is a practical one and, whenever he has recourse to numerical work, he chooses his examples with considerable skill.

The book begins with a definition of terms and an explanation of the way in which various kinds of differences, and particularly divided differences, are used and then gives interpolation formulae including Bessel's, Stirling's, Everett's and other well known expressions. After sections which deal with Factorial Coefficients and Numerical Differentiation we come to a helpful discussion of the application of the method of Finite Differences to the Construction of Tables and then, after having the pleasure of finding the so-called Inverse Interpolation dealt with logically, we reach summation, repeated summation and the formulae used for summation (Gauss, Euler, Lubbock, etc.).

In the next section on Mechanical Quadrature some care has to be exercised by the reader because in two successive formulae which relate to the "open" and "closed" types respectively the same notation is used to imply different ordinates. Bearing in mind the author's development of the formulae this is not altogether unnatural, but we think it would have been preferable to have used a distinguishing notation in the final formulae or to have shown more clearly exactly which ordinates have to be used in each case.

The numerical integration of differential equations is discussed and, at the end of the book, we find sections on Interpolation with Several Variables and Mechanical Cubature. The author has, whenever possible, given a Remainder Term with the Interpolation or Quadrature formulae and this, combined with the comments which he makes from time to time on the use of the formulae, enables the reader to form an idea of their limitations or of the error involved in their use. Incidentally it has given Professor Steffensen the opportunity of bringing into book form much of his own work on the subject which was formerly available only in various Scandinavian Journals.

Although there are many complicated mathematical expressions throughout the book and some difficult type-setting we have noticed few misprints. We must confess however that we do not think the type used for the mathematical work is as pleasing as that of the best English mathematical books, and the paper of which the book is made, being rather soft in surface, does not give the smaller types the best chance and will not, we think, stand hard wear.

Professor Steffensen says in his preface that he wishes it to be understood that the book is meant as a textbook and not as a handbook or encyclopaedia. In our view it will prove to be valuable both as a textbook and a book of reference: taking it as a whole, we have not seen a better book on the subject.

W. P. E.

CAMBRIDGE: PRINTED BY
W. LEWIS, M.A.
AT THE UNIVERSITY PRESS